On Generalizing Collective Spatial Keyword Queries

Chan, H. K-H., Long, C., & Wong, R. C-W. (2018). On Generalizing Collective Spatial Keyword Queries. *IEEE Transactions on Knowledge and Data Engineering*, *30*(9), 1712-1726.
https://doi.org/10.1109/TKDE.2018.2800746

Published in:
IEEE Transactions on Knowledge and Data Engineering

Document Version:
Peer reviewed version

Queen's University Belfast - Research Portal:
Link to publication record in Queen's University Belfast Research Portal

On Generalizing Collective Spatial Keyword Queries

Harry Kai-Ho Chan, Cheng Long, and Raymond Chi-Wing Wong

Abstract—With the proliferation of spatial-textual data such as location-based services and geo-tagged websites, spatial keyword queries are ubiquitous in real life. One example of spatial-keyword query is the so-called *collective spatial keyword query* (CoSKQ) which is to find for a given query consisting a query location and several query keywords a set of objects which *covers* the query keywords collectively and has the smallest *cost* wrt the query location. In the literature, many different functions were proposed for defining the *cost* and correspondingly, many different approaches were developed for the CoSKQ problem. In this paper, we study the CoSKQ problem systematically by proposing *a unified cost function* and *a unified approach* for the CoSKQ problem (with the unified cost function). The unified cost function includes all existing cost functions as special cases and the unified approach solves the CoSKQ problem with the unified cost function in a unified way. Experiments were conducted on both real and synthetic datasets which verified our proposed approach.

Index Terms—Spatial keyword queries, unified framework

1 Introduction

Nowadays , geo-textual data which refers to data with both spatial and textual information is ubiquitous. Some examples of geo-textual data include the spatial points of interest (POI) with textual description (e.g., restaurants, cinema, tourist attractions, and hotels), geo-tagged web objects (e.g., webpages and photos at Flickr), and also geo-social networking data (e.g., users of FourSquare have their check-in histories which are spatial and also profiles which are textual).

One application based on geo-textual data is to search a set of (geo-textual) objects wrt a query consisting of a query location (e.g., the location one is located at) and some textual information (e.g., some keywords expressing the targets one wants to search) such that the objects have their textual information *matching* the query keywords and their locations close to the query location. One scenario of this application is that a tourist wants to find several POIs such that s/he could do sight-seeing, shopping and dining and the POIs are close to the hotel. In this case, the user can set the query location to the hotel location and the query keywords to be "attractions", "shopping" and "restaurant" to search for a set of POIs. Another scenario is that a manager wants to set up a project consortium of partners close to each other such that they together offer the capabilities required for successful execution of the whole project. In this case, the user can issue the query with his/her location as the query location and the required skills for the partners as the query keywords to find a group of people.

The above applications were captured by the so-called *Collective Spatial Keyword Query* (CoSKQ) [3], [17], [2] in the literature. Let $\mathcal{O}$ be a set of objects, where each object $o \in \mathcal{O}$ is associated with a spatial location, denoted by $o.\lambda$, and a set of keywords, denoted by $o.\psi$. Given a query q with a location $q.\lambda$ and a set of keywords $q.\psi$, the CoSKQ problem is to find a set S of objects such that S covers $q.\psi$, i.e., $q.\psi \subseteq \cup_{o \in S} o.\psi$, and the $cost$ of S, denoted by $cost(S)$, is minimized.

In the literature, many different cost functions have been proposed for $cost(S)$ in the CoSKQ problem, and these cost functions are applicable in different scenarios in addition to the above examples. For the CoSKQ problem with each particular cost function, at least one approach has been designed, which we briefly review as follows.

Different cost functions. Five different cost functions have been proposed for the CoSKQ problem, namely, $cost_{Sum}$ [3], $cost_{MaxMax}$ [3], $cost_{MaxMax2}$ [17], $cost_{MinMax}$ [2] and $cost_{SumMax}$ [2]. For example, $cost_{Sum}(S)$ defines the cost to the summation of the distances from the query location to the objects in S, and $cost_{MaxMax}(S)$ defines the cost to a linear combination of the maximum distance between the query location and an object in S and the maximum pairwise distance among the objects in S. The definitions of the rest of cost functions would be introduced later. Each cost function has its own semantic meaning and depending on the application scenario, an appropriate cost function is used.

Different approaches. For the CoSKQ problem with each of these existing cost functions, which was proved to be NP-hard, at least one solution (including an exact algorithm and an approximate algorithm) was developed, and these solutions usually differ from one another. For example, the exact algorithm for the CoSKQ problem with $cost_{Sum}$ is a dynamic programming algorithm [3], while that for the one with $cost_{MaxMax}$ is a branch-and-bound algorithm [3]. Usually, an existing algorithm for the CoSKQ problem with a particular cost function cannot be used to solve that with another cost function.

In this paper, we study the CoSKQ problem systematically by proposing *a unified cost function* and *a unified approach* for the CoSKQ problem (with the unified cost function).

Without the unified approach, we need to handle different cost functions by different algorithms, which increases the difficulty

- *H.K.-H. Chan and R.C.-W. Wong are with the Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong. E-mail: {khchanak, raywong}@cse.ust.hk*
- *C. Long is with the School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, BT7 1NN, Northern Ireland, United Kingdom. E-mail: cheng.long@qub.ac.uk*

for CoSKQ to be used in practice. Also, when researchers work on improving the performance of an algorithm, only the corresponding cost function is benefited. Although sometimes it is possible that one algorithm originally designed for one cost function can be adapted for another cost function, the performance of the adapted algorithm is not satisfactory. A better idea is to have a unified cost function and a unified approach, where the unified cost function captures all known cost functions and some other cost functions which are not known before but useful.

Specifically, the main contribution is summarized as follows.

A unified cost function. We propose a unified cost function $cost_{unified}$ which expresses all existing cost functions and a few new cost functions that have not been studied before. The core idea of $cost_{unified}$ is that first two distance components, namely the *query-object distance component* and the *object-object distance component*, are defined, where the former is based on the distances between the query location and those of the objects and the latter is based on the pairwise distances among the set of objects and then $cost_{unified}$ is defined based on the two distance components carefully such that all existing cost functions are captured (Note that this is possible since all ingredients of defining a cost function are distances between the query location and and those distances among objects which are captured by the two components.).

A unified approach. We design a unified approach, which consists of one exact algorithm and one approximate algorithm, for the CoSKQ problem with the unified cost function. For the CoSKQ problem with the cost function instantiated to those existing cost functions, which have been proved to be NP-hard, our exact algorithm is superior over the state-of-the-arts in that it not only has a *unified* procedure, but also runs faster under *all* settings for some cost functions (e.g., $cost_{MinMax}$ and $cost_{MinMax2}$) and under the *majority* of settings for the other cost functions, and our approximate algorithm is always among those algorithms which give the best approximation ratios and runs faster than those algorithms which give similar approximation ratios. For the CoSKQ problem with the cost function instantiated to those new cost functions that have not been studied before, our exact algorithm runs reasonably fast and our approximate algorithm provides certain approximation ratios.

Besides, we conducted extensive experiments based on both real and synthetic datasets which verified our unified approach.

The rest of this paper is organized as follows. Section 2 gives the related work. Section 3 introduces the unified cost function and Section 4 presents the unified approach for CoSKQ. Section 5 gives the empirical study and Section 6 concludes the paper.

2 Related Work

Many existing studies on spatial keyword queries focus on retrieving a *single object* that is close to the query location and relevant to the query keywords.

A *boolean kNN query* [12], [5], [24], [30], [27] finds a list of k objects each covering all specified query keywords. The objects in the list are ranked based on their spatial proximity to the query location.

A *top-k kNN query* [8], [18], [15], [19], [20], [9], [25] adopts the ranking function considering both the spatial proximity and the textual relevance of the objects and returns top-k objects based on the ranking function. This type of queries has been studied on Euclidean space [8], [18], [15], road network databases [19], trajectory databases [20], [9] and moving object databases [25]. Usually, the methods for this kind of queries adopt an index structure called the *IR-tree* [8], [23] capturing both the spatial proximity and the textual information of the objects to speed up the keyword-based nearest neighbor (NN) queries and range queries. In this paper, we also adopt the *IR-tree* for keyword-based NN queries and range queries.

Some other studies on spatial keyword queries focus on finding an *object set* as a solution. Among them, some [3], [17], [2] studied the *collective spatial keyword queries* (CoSKQ). Cao et al. [3], [2] proposed four cost functions, namely $cost_{Sum}$, $cost_{MaxMax}$, $cost_{MinMax}$ and $cost_{SumMax}$, and developed algorithms for the CoSKQ problem with the first three cost functions, leaving that with the fourth cost function, i.e., $cost_{SumMax}$, as future work. Besides, they studied two variations of CoSKQ, namely top-k CoSKQ and weighted CoSKQ, in [2]. Long et al. [17] proposed exact and approximate algorithms for the CoSKQ problem with $cost_{MaxMax}$ and also that with a new cost function $cost_{MaxMax2}$. The details of these cost functions are described in Section 3. In this paper, we also study the CoSKQ problem. Specifically, we propose a *unified* cost function which include all existing cost functions as special cases and based on the unified cost function, we design a *unified* approach, consisting of an exact algorithm and an approximate algorithm.

Another query that is similar to the CoSKQ problem is the *mCK query* [28], [29], [14] which takes a set of m keywords as input and finds m objects with the minimum *diameter* that cover the m keywords specified in the query. In the existing studies of mCK queries, it is usually assumed that each object contains a single keyword. There are some variants of the mCK query, including the *SK-COVER* [7] and the *BKC query* [10]. These queries are similar to the CoSKQ problem in that they also return an object set that covers the query keywords, but they only take a set of keywords as input. In contrast, the CoSKQ problem studied in this paper takes both a set of keywords and a spatial location as inputs.

Skovsgaard et al. [21] proposed a query to find top-k groups of objects with the ranking function considering the spatial proximity and textual relevance of the groups. Liu et al. proposed the *clue-based spatio-textual query* [16] which takes a set of keywords and a clue as inputs, and returns k objects with highest similarities against the clue.

There are also some studies [13], [22] on spatial keyword queries which find an object set in the road network, some [6] which find an object set with the scoring function considering an inherent cost in each object, some [4], [11] which find a *region* as a solution and some [1], [26] which find a *route* as a solution.

3 A Unified Cost Function

Let $\mathcal{O}$ be a set of objects, where each object $o \in \mathcal{O}$ is associated with a spatial location, denoted by $o.\lambda$, and a set of keywords, denoted by $o.\psi$. Given two objects o_1 and o_2, we denote by $d(o_1, o_2)$ the Euclidean distance between $o_1.\lambda$ and $o_2.\lambda$.

(1) Problem definition. A *collective spatial keyword query* (CoSKQ) [3] is defined as follows.

Problem 1 (CoSKQ [3]). Given a query q with a location $q.\lambda$ and a set of keywords $q.\psi$, **the CoSKQ problem** is to find a set S of objects such that S covers $q.\psi$, i.e., $q.\psi \subseteq \cup_{o \in S} o.\psi$, and the $cost$ of S, denoted by $cost(S)$, is minimized. □

(2) Existing cost functions. To the best of our knowledge, five cost functions have been proposed for defining $cost(\cdot)$ in

	Parameter			$cost_{unified}(S\|\alpha,\phi_1,\phi_2)$	Existing/New
	$\alpha \in (0,1]$	$\phi_1 \in \{1,\infty,-\infty\}$	$\phi_2 \in \{1,\infty\}$		
a	0.5*	1	1	$\sum_{o\in S} d(o,q) + \max_{o_1,o_2\in S} d(o_1,o_2)$	$cost_{SumMax}$ [2]
b	0.5*	1	∞	$\max\{\sum_{o\in S} d(o,q), \max_{o_1,o_2\in S} d(o_1,o_2)\}$	$cost_{SumMax2}$ (New)
c	0.5*	∞	1	$\max_{o\in S} d(o,q) + \max_{o_1,o_2\in S} d(o_1,o_2)$	$cost_{MaxMax}$ [3], [17], [2]
d	0.5*	∞	∞	$\max\{\max_{o\in S} d(o,q), \max_{o_1,o_2\in S} d(o_1,o_2)\}$	$cost_{MaxMax2}$ [17]
e	0.5*	$-\infty$	1	$\min_{o\in S} d(o,q) + \max_{o_1,o_2\in S} d(o_1,o_2)$	$cost_{MinMax}$ [2]
f	0.5*	$-\infty$	∞	$\max\{\min_{o\in S} d(o,q), \max_{o_1,o_2\in S} d(o_1,o_2)\}$	$cost_{MinMax2}$ (New)
g	1	1	-	$\sum_{o\in S} d(o,q)$	$cost_{Sum}$ [3], [2]
h	1	∞	-	$\max_{o\in S} d(o,q)$	$cost_{Max}$ (New)
i	1	$-\infty$	-	$\min_{o\in S} d(o,q)$	$cost_{Min}$ (New)

* Following the existing studies, $\alpha = 0.5$ is used to illustrate the case of $\alpha \in (0,1)$ for simplicity

TABLE 1: $cost_{unified}$ under different parameter settings

the CoSKQ problem, namely $cost_{Sum}$ [3], $cost_{SumMax}$ [2], $cost_{MaxMax}$ [3], $cost_{MaxMax2}$ [17], and $cost_{MinMax}$ [2]. Specifically, these cost functions are defined as follows.

1) $cost_{Sum}$. $cost_{Sum}(S)$ defines the cost to be the summation of the distances from the query location to the objects in S, i.e., $cost_{Sum}(S) = \sum_{o\in S} d(o,q)$.
2) $cost_{SumMax}$. $cost_{SumMax}(S)$ defines the cost to be a linear combination of the summation of distances from the query location to the objects in S and the maximum pairwise distance among the objects in S, i.e., $cost_{SumMax}(S) = \alpha \cdot \sum_{o\in S} d(o,q) + (1-\alpha) \cdot \max_{o_1,o_2\in S} d(o_1,o_2)$, where α represents a real number in $(0,1]$.
3) $cost_{MaxMax}$. $cost_{MaxMax}(S)$ defines the cost to be a linear combination of the maximum distance between the query location and an object in S and the maximum pairwise distance among the objects in S, i.e., $cost_{MaxMax}(S) = \alpha \cdot \max_{o\in S} d(o,q) + (1-\alpha) \cdot \max_{o_1,o_2\in S} d(o_1,o_2)$, where α represents a real number in $(0,1]$.
4) $cost_{MaxMax2}$. $cost_{MaxMax2}(S)$ defines the cost to be the larger one of the maximum distance between the query location and an object in S and the maximum pairwise distance among the objects in S, i.e., $cost_{MaxMax2}(S) = \max\{\max_{o\in S} d(o,q), \max_{o_1,o_2\in S} d(o_1,o_2)\}$.
5) $cost_{MinMax}$. $cost_{MinMax}(S)$ defines the cost to be a linear combination of the minimum distance between the query location and an object in S and the maximum pairwise distance among the objects in S, i.e., $cost_{MinMax}(S) = \alpha \cdot \min_{o\in S} d(o,q) + (1-\alpha) \cdot \max_{o_1,o_2\in S} d(o_1,o_2)$, where α represents a real number in $(0,1]$.

(3) A unified cost function $cost_{unified}$. In this paper, we propose a *unified* cost function $cost_{unified}$ which could be instantiated to many different cost functions including all those five existing ones. Before we give the exact definition of $cost_{unified}$, we first introduce a distance component used for defining $cost_{unified}$, namely the *query-object distance component*. It is defined based on the distances between the query location and the objects in S. Specifically, we denote it by $D_{q,o}(S|\phi_1)$ and define it as follows.

$$D_{q,o}(S|\phi_1) = [\sum_{o\in S}(d(o,q))^{\phi_1}]^{\frac{1}{\phi_1}}$$

where $\phi_1 \in \{1,\infty,-\infty\}$ is a user parameter. Depending on the setting of ϕ_1, $D_{q,o}(S|\phi_1)$ corresponds to the summation, the maximum, or the minimum of the distances from the query location to the objects in S. Specifically,

$$D_{q,o}(S|\phi_1) = \begin{cases} \sum_{o\in S} d(o,q), & \text{if } \phi_1 = 1 \\ \max_{o\in S} d(o,q), & \text{if } \phi_1 = \infty \\ \min_{o\in S} d(o,q), & \text{if } \phi_1 = -\infty \end{cases}$$

With the distance component defined, we are ready to introduce the unified cost function $cost_{unified}$. Specifically, we define $cost_{unified}$ as follows.

$$\begin{aligned} &cost_{unified}(S|\alpha,\phi_1,\phi_2) \\ &= \{[\alpha \cdot D_{q,o}(S|\phi_1)]^{\phi_2} + [(1-\alpha) \max_{o_1,o_2\in S} d(o_1,o_2)]^{\phi_2}\}^{\frac{1}{\phi_2}} \end{aligned} \quad (1)$$

where $\alpha \in (0,1]$[1], $\phi_1 \in \{1,\infty,-\infty\}$ and $\phi_2 \in \{1,\infty\}$ are user parameters. In the following, we write $cost_{unified}(S|\alpha,\phi_1,\phi_2)$ simply as $cost(S)$ when there is no ambiguity.

Same as [3], [17], [2], for ease of exposition, we use $\alpha = 0.5$ to illustrate the case of $\alpha \in (0,1)$. In this case, we can safely assume that

$$\begin{aligned} &cost_{unified}(S|0.5,\phi_1,\phi_2) \\ &= \{[D_{q,o}(S|\phi_1)]^{\phi_2} + [\max_{o_1,o_2\in S} d(o_1,o_2)]^{\phi_2}\}^{\frac{1}{\phi_2}} \end{aligned} \quad (2)$$

Under some settings of α, ϕ_1 and ϕ_2, $cost_{unified}$ corresponds to one of the aforementioned existing cost functions (as shown in Table 1). For example, when $\alpha = 1$ and $\phi_1 = 1$ (regardless of the settings of ϕ_2), $cost_{unified}(S)$ corresponds to $cost_{Sum}(S)$ since

$$\begin{aligned} cost_{unified}(S) &= \{[D_{q,o}(S|1)]^{\phi_2}\}^{\frac{1}{\phi_2}} = D_{q,o}(S|1) \\ &= \sum_{o\in S} d(o,q) = cost_{Sum}(S) \end{aligned}$$

and similarly, when $\alpha \in (0,1]$, $\phi_1 = \infty$ and $\phi_2 = 1$, $cost_{unified}(S)$ corresponds to $cost_{MaxMax}(S)$.

Under some other settings of α, ϕ_1 and ϕ_2, $cost_{unified}$ corresponds to a new cost function that has not been studied before. For example, when $\alpha = 0.5$, $\phi_1 = 1$, and $\phi_2 = \infty$, we have

$$\begin{aligned} cost_{unified}(S) &= \{[0.5 \cdot D_{q,o}(S|1)]^{\infty} + [0.5 \cdot \max_{o_1,o_2\in S} d(o_1,o_2)]^{\infty}\}^{\frac{1}{\infty}} \\ &= 0.5 \max\{\sum_{o\in S} d(o,q), \max_{o_1,o_2\in S} d(o_1,o_2)\} \end{aligned}$$

where we denote $\max\{\sum_{o\in S} d(o,q), \max_{o_1,o_2\in S} d(o_1,o_2)\}$ by $cost_{SumMax2}(S)$.

The instantiations of $cost_{unified}$ depending on different parameter settings are shown in Table 1. In the following, we introduce those instantiations that are new.

1. In the setting of $\alpha = 0$, the query location has no contribution to the cost. Thus, we do not consider this setting.

1) (row b) $cost_{SumMax2}$. The functionality of this cost function is equivalent to that of the cost function $cost_{Sum}$ (please see Appendix A for details), and thus we focus on $cost_{Sum}$ in this paper.
2) (row f) $cost_{MinMax2}$. It essentially captures the maximum among two distances, namely the distance between the query location $q.\lambda$ and its nearest object in S and the distance between the two farthest objects in S. A common practice for an individual to explore the objects returned is to visit the object which is the nearest from the query location and explore the others, and thus this cost function is useful when people want to get at their first stop (i.e., the nearest object) fastly (this is captured by the query-object distance component) and explore the objects within a small region (this is captured by the farthest pairwise distance of the objects). Compared to the existing cost function $cost_{MinMax}$, $cost_{MinMax2}$ has an advantage that it requires no parameter of α.
3) (row h) $cost_{Max}$. It uses the maximum distance between the query location $q.\lambda$ and an object in S. This cost function can be used to find the feasible set with the closet farthest object among all feasible sets. This cost function is suitable for the scenarios where a user visits one object a time, starting from the query location each time, and wants the worst-case cost as small as possible.
4) (row i) $cost_{Min}$. It uses the distance between the query location $q.\lambda$ and its nearest object in S only, which is of no interest in practice since it put no penalty on those objects that are far away from the query location, e.g., the whole set of objects corresponds to a trivial solution for the CoSKQ problem with $cost_{Min}$. Therefore, we ignore this instantiation of $cost_{unified}$.

(4) Intractability results. It is known that the CoSKQ problem with an existing cost function adopted is NP-hard [3], [17], [2]. That is, the CoSKQ problem is NP-hard under the parameter settings such that $cost_{unified}$ corresponds to an existing cost function. In this paper, we study the intractability of the CoSKQ problem with all possible parameter settings of α, ϕ_1 and ϕ_2 for $cost_{unified}$. Specifically, we have the following result.

Theorem 1 (Intractability). The CoSKQ problem is NP-hard with all possible parameter settings of α, ϕ_1 and ϕ_2 except for the setting of $\alpha = 1, \phi_1 \in \{\infty, -\infty\}$. □

Proof. See Appendix B. □

(5) Existing Algorithms. For the CoSKQ problem with each of the existing cost functions, solution (including an exact algorithm and an approximate algorithm) was developed, and these solutions usually differ from one another. Specifically, we review the algorithms of some existing cost functions and solutions as follows.

1) $cost_{Sum}$. The exact algorithm for CoSKQ problem with $cost_{Sum}$ is a dynamic programming algorithm, while the approximate algorithm is a greedy algorithm transformed from that of the Weighted Set Cover problem [3], [2]
2) $cost_{SumMax}$. No solution is available in the literature for solving CoSKQ with $cost_{SumMax}$. This cost function is proposed in [2], but the corresponding solution is left for their future work.
3) $cost_{MaxMax}$. Several algorithms were proposed for CoSKQ problem with $cost_{MaxMax}$. One of the exact algorithms is a branch-and-bound algorithm [3], while another one is based on a distance owner-driven approach [17]. One of the approximate algorithms picks the nearest neighbor set [3], [2], while two other approximate algorithms search for feasible sets in an iterative manner [3], [17], [2].

Usually, an existing algorithm for the CoSKQ problem with a particular cost function cannot be used to solve that with another cost function. In the following section, we introduce out unified approach for the CoSKQ problem with the unified cost function.

4 A Unified Approach

In this section, we introduce our unified approach which consists of one exact algorithm called *Unified-E* (Section 4.1) and one approximate algorithm called *Unified-A* (Section 4.2). While the unified cost function combines existing ones, our unified approach is not one which simply combine existing approaches. In fact, both the exact algorithm and approximate algorithm proposed in this paper are clean and elegant while existing approaches have quite different structures.

Before presenting the algorithms, we first give some definitions as follows. Given a query q and an object o in $\mathcal{O}$, we say o is a **relevant object** if $o.\psi \cap q.\psi \neq \emptyset$. We denote $\mathcal{O}_q$ to be the set of all relevant objects. Given a set S of objects, S is said to be a **feasible set** if S covers $q.\psi$ (i.e. $q.\psi \subseteq \cup_{o \in S} o.\psi$). Note that the CoSKQ problem is to find a feasible set with the smallest cost.

Given a non-negative real number r, we denote the circle centered at $q.\lambda$ with radius r by $C(q, r)$. Similarly, the circle centered at $o.\lambda$ with radius r is denoted by $C(o, r)$.

Let q be a query and S be a feasible set. We say that an object $o \in S$ is a **query-object distance contributor** wrt S if $d(o, q)$ *contributes in* $D_{q,o}(S|\phi_1)$. Specifically, we have the following three cases according to the value of ϕ_1.

- In the case of $\phi_1 = 1$ where $D_{q,o}(S|\phi_1) = \sum_{o \in S} d(o, q)$, each object in S is a query-object distance contributor wrt S;
- In the case of $\phi_1 = \infty$ where $D_{q,o}(S|\phi_1) = \max_{o \in S} d(o, q)$, only those objects in S which have the maximum distance from q are the query-object distance contributors wrt S;
- In the case of $\phi_1 = -\infty$ where $D_{q,o}(S|\phi_1) = \min_{o \in S} d(o, q)$, only those objects in S which have the minimum distance from q are the query-object distance contributors wrt S.

Then, we define the **key query-object distance contributor** wrt S to the object with the greatest distance from q among all query-object distance contributors wrt S. The concept of "key query-object distance contributor" is inspired by the concept of "query distance owner" proposed in [17], and the concept of "key query-object distance contributor" is more general in the sense that a query distance owner corresponds to a key query distance contributor in the case of $\phi_1 = \infty$ but not in other cases.

Let S be a set of objects and o_i and o_j are two objects in S. We say that o_i and o_j are **object-object distance contributors** wrt S if $d(o_i, o_j)$ *contribute in* $\max_{o,o' \in S} d(o, o')$, i.e, $(o_i, o_j) = \arg\max_{o,o' \in S} d(o, o')$.

Given a query q and a keyword t, the **t-keyword nearest neighbor** of q, denoted by $NN(q, t)$, it defined to be the nearest neighbor (NN) of q containing keyword t. Similarly, $NN(o, t)$

is defined to be the NN of o containing keyword t. Besides, we define the **nearest neighbor set** of q, denoted by $N(q)$ to be the set containing q's t-keyword nearest neighbor for each $t \in q.\psi$, i.e., $N(q) = \cup_{t \in q.\psi} NN(q,t)$. Note that $N(q)$ is a feasible set.

4.1 An Exact Algorithm

The idea of *Unified-E* is to iterate through the object-object distance contributors and search for the best feasible set S' in each iteration. This allows CoSKQ with different cost functions to be executed efficiently. Note that each existing algorithm [3], [17], [2] is designed for a specific cost function and they cannot be used to answer CoSKQ with different cost functions.

Specifically, *Unified-E* adopts the following search strategy.

- Step 1 (Object-Object Distance Contributors Finding): Select two objects to be the object-object distance contributors wrt the set S' to be constructed;
- Step 2 (Key Query-Object Distance Contributor Finding): Select an object to be the key query-object distance contributor wrt the set S' to be constructed;
- Step 3 (Best Feasible Set Construction): Construct the set S' (which has o_i, o_j as the object-object distance contributors and o_m as the key query-object distance contributor), and update the current best solution $curSet$ with S' if $cost(S') < curCost$, where $curCost$ is the cost of $curSet$;
- Step 4 (Iterative Step): Repeat Step 1 to Step 3 until all possible object-object distance contributors and key query-object distance contributors are iterated.

The above search strategy makes quite effective pruning possible at both Step 1 and Step 2.

Pruning at Step 1. The major idea is that not each relevant objects pair is necessary to be considered as a object-object distance contributor wrt S' to be constructed. First, only the relevant objects in $R_S = C(q, r_1)$ need to be considered, where r_1 is the radius of the region that depends on the parameter setting, as shown in Table 2. It can be proved that if S' contains an object o such that $d(o,q) > r_1$, S' cannot be the optimal solution. Second, we can maintain a lower bound d_{LB} and an upper bound d_{UB} of the distance between the object-object distance contributors for pruning. For example, all those relevant objects pairs (o_i, o_j) with $d(o_i, o_j) > curCost$ (this is because in this case, all those feasible sets S' with (o_i, o_j) as the object-object distance contributor have the cost larger than that of the current best solution, i.e., the best-known cost) could be pruned, i.e., $curCost$ is used as an upper bound. Furthermore, it could be verified easily that when $\phi_1 \in \{1, \infty\}$, all those relevant object pairs (o_i, o_j) with $d(o_i, o_j) < \max_{o \in N(q)} d(o,q) - \min\{d(o_i,q), d(o_j,q)\}$ could be pruned, i.e., $\max_{o \in N(q)} d(o,q) - \min\{d(o_i,q), d(o_j,q)\}$ is used as a lower bound. The details of d_{LB} and d_{UB} for different parameter settings are presented in Table 2. Specifically, we have the following lemma.

Lemma 1. Let o_i and o_j be the object-object distance contributors of the set S to be constructed. For $cost_{unified}$ with different parameter settings, $d(o_i, o_j)$ can be lower bounded by d_{LB} and upper bounded by d_{UB}, as shown in Table 2. □

Proof. Let o_m be the key query-object distance contributor of S. The proof of d_{LB} is shown as follows. When $\phi_1 \in \{1, \infty\}$, $d(o_i, o_j) \geq d(o_i, o_m)$ and $d(o_i, o_j) \geq d(o_j, o_m)$. Besides, we know that $d(o_i, o_m) + d(o_i, q) \geq d(o_m, q)$ by triangle inequality. Similarly, we know that $d(o_j, o_m) + d(o_j, q) \geq d(o_m, q)$. Since S is feasible, $d(o_m, q) \geq d_f$. Therefore, we have $d(o_i, o_j) \geq d(o_m, q) - min\{d(o_i, q), d(o_j, q)\} \geq d_f - min\{d(o_i, q), d(o_j, q)\} = d_{LB}$. When $\phi_1 = -\infty$, we have $d(o_m, q) + d(o_i, o_j) \geq d_f$ because S is feasible. Also, $d(o_i, q) \geq d(o_m, q)$ and $d(o_j, q) \geq d(o_m, q)$ because o_m is the object closet to q. Therefore, we have $d(o_i, o_j) \geq d_f - d(o_m, q) \geq d_f - min\{d(o_i, q), d(o_j, q)\} = d_{LB}$.

The proof of d_{UB} is shown as follows. When $\alpha = 0.5, \phi_1 = 1$ and $\phi_2 = 1$ ($cost_{SumMax}$), $cost(S) \geq d(o_i, q) + d(o_j, q) + d(o_i, o_j)$ and $d(o_i, q) + d(o_j, q) \geq d(o_i, o_j)$ by triangle inequality. If $d(o_i, o_j) \geq curCost/2$, we have $cost(S) \geq 2d(o_i, o_j) \geq curCost$, which means S cannot contribute to a better solution and can be pruned. When $\alpha = 0.5, \phi_1 = \infty$ and $\phi_2 = 1$ ($cost_{MaxMax}$), $cost(S) = d(o_m, q) + d(o_i, o_j)$ and $d(o_m, q) \geq d_f$ since S is a feasible set. If $d(o_i, o_j) \geq curCost - d_f$, we have $cost(S) \geq curCost$ and thus S can be pruned. For the other parameter settings, it is easy to see that if S contain an object o with $d(o, q) \geq curCost$, $cost(S) \geq curCost$. □

Third, given a set having o_i and o_j as the object-object distance contributors, we can compute the lower bound of cost of the set, denoted by $cost(\{o_i, o_j\})_{LB}$, and thus we can prune all those object pairs with $cost(\{o_i, o_j\})_{LB} > curCost$. The details of $cost(\{o_i, o_j\})_{LB}$ for different parameter settings are presented in Table 2. Specifically, we have the following lemma.

Lemma 2. Let o_i and o_j be the object-object distance contributors of the set S to be constructed. For $cost_{unified}$ with different parameter settings, $cost(S)$ can be lower bounded by $cost(\{o_i, o_j\})_{LB}$, as shown in Table 2. □

Proof. Let o_m be the key query-object distance contributor of S. When $\phi_1 = 1$, it is obvious that $cost(S) \geq cost(\{o_i, o_j\})_{LB}$.

When $\phi_1 = \infty$, $d(o_m, q) \geq \max\{d(o_i, q), d(o_j, q)\}$. Since S is a feasible set, $d(o_m, q) \geq d_f$. Thus, $cost(S) \geq d(o_i, o_j) + \max\{d(o_i, q), d(o_j, q), d_f\}$ when $\phi_2 = 1$ ($cost_{MaxMax}$) and $cost(S) \geq \max\{d(o_i, o_j), d(o_i, q), d(o_j, q), d_f\}$ when $\phi_2 = \infty$ ($cost_{MaxMax2}$).

When $\phi_1 = -\infty$ and $\phi_2 = 1$ ($cost_{MinMax}$), we know that $cost(S) \geq d(o_i, o_j)$. Also we have $cost(S) \geq d(o_m, q) + d(o_i, o_m) \geq d(o_i, q)$ by triangle inequality. Similarly, we have $cost(S) \geq d(o_m, q) + d(o_j, o_m) \geq d(o_j, q)$. Therefore, $cost(S) \geq \max\{d(o_i, o_j), d(o_i, q), d(o_j, q)\}$.

When $\phi_1 = -\infty$ and $\phi_2 = \infty$ ($cost_{MinMax2}$), we know that $cost(S) \geq d(o_i, o_j)$. Also $d(o_m, q) \geq d(o_i, q) - d(o_i, o_j)$ because o_m must be located in the region of $C(o_i, d(o_i, o_j))$. Similarly, $d(o_m, q) \geq d(o_j, q) - d(o_i, o_j)$. Therefore, $cost(S) \geq \max\{d(o_i, o_j), \max\{d(o_i, q), d(o_j, q)\} - d(o_i, o_j)\}$. □

Pruning at Step 2. Note that only the objects in $C(o_i, d(o_i, o_j)) \cap C(o_j, d(o_i, o_j))$ need to be considered as key query-object distance contributors for constructing S'. The major idea of the pruning is that not all possible objects in the region are necessary to be considered. Specifically, we can maintain a lower bound r_{LB} and an upper bound r_{UB} of the distance between the key query-object distance contributors and query. For example, in the case that $\phi_1 = 1$, all those relevant objects o with $d(o, q) < \max\{d(o_i, q), d(o_j, q)\}$ could be safely pruned (this is because such object o can not be the key query-object distance contributor wrt S'), i.e., $\max\{d(o_i, q), d(o_j, q)\}$ is used as lower bound. Figure 1(a) shows the region for the objects to be considered as the key query-object distance contributor. In the

Cost function	Parameter			r_1	d_{LB}	d_{UB}	$cost(\{o_i, o_j\})_{LB}$
	α	ϕ_1	ϕ_2				
$cost_{SumMax}$	0.5	1	1	$curCost$	$d_f - \min\{d(o_i,q), d(o_j,q)\}$	$curCost/2$	$d(o_i,o_j) + d(o_i,q) + d(o_j,q)$
$cost_{MaxMax}$	0.5	∞	1	$curCost$		$curCost - d_f$	$d(o_i,o_j) + \max\{d(o_i,q), d(o_j,q), d_f\}$
$cost_{MaxMax2}$	0.5	∞	∞	$curCost$		$curCost$	$\max\{d(o_i,o_j), d(o_i,q), d(o_j,q), d_f\}$
$cost_{MinMax}$	0.5	$-\infty$	1	$curCost$		$curCost$	$\max\{d(o_i,o_j), d(o_i,q), d(o_j,q), d_f\}$
$cost_{MinMax2}$	0.5	$-\infty$	∞	$2 \cdot curCost$		$curCost$	$\max\{d(o_i,o_j), \max\{d(o_i,q), d(o_j,q)\} - d(o_i,o_j)\}$
$cost_{Sum}$	1	1	-	$curCost$		$curCost$	$d(o_i,q) + d(o_j,q)$

$d_f = \max_{o \in N(q)} d(o,q)$

TABLE 2: Lower and upper bounds used in Step 1 of Unified-E

Cost function	Parameter			r_{LB}	r_{UB}
	α	ϕ_1	ϕ_2		
$cost_{SumMax}$	0.5	1	1	$\max\{d(o_i,q), d(o_j,q), d_f\}$	$curCost - d(o_i,o_j)$
$cost_{MaxMax}$	0.5	∞	1	$\max\{d(o_i,q), d(o_j,q), d_f\}$	$curCost - d(o_i,o_j)$
$cost_{MaxMax2}$	0.5	∞	∞	$\max\{d(o_i,o_j), d_f\}$	$curCost$
$cost_{MinMax}$	0.5	$-\infty$	1	$d_f - d(o_i,o_j)$	$\min\{curCost - d(o_i,o_j), \min\{d(o_i,q), d(o_j,q)\}\}$
$cost_{MinMax2}$	0.5	$-\infty$	∞	$d_f - d(o_i,o_j)$	$\min\{d(o_i,q), d(o_j,q)\}$
$cost_{Sum}$	1	1	-	$\max\{d(o_i,q), d(o_j,q), d_f\}$	$curCost - d(o_i,q) - d(o_j,q)$

$d_f = \max_{o \in N(q)} d(o,q)$

TABLE 3: Lower and upper bounds used in Step 2 of Unified-E

case that $\phi_1 = -\infty$, similarly, all those relevant objects o with $d(o,q) > \min\{d(o_i,q), d(o_j,q)\}$ could be safely pruned i.e., $\min\{d(o_i,q), d(o_j,q)\}$ is used as upper bound. Also, all those relevant objects o with $d(o,q) < d_f - d(o_i,o_j)$ could be safely pruned, where $d_f = \max_{o \in N(q)} d(o,q)$ (this is because all those feasible sets S' with o as the key query-object distance contributor have $\max_{o_1,o_2 \in S'} d(o_1,o_2)$ larger than $d(o_i,o_j)$), i.e., $d_f - d(o_i,o_j)$ is used as an lower bound. Figure 1(b) shows the region for the objects to be considered as the key query-object distance contributor. The details of r_{LB} and r_{UB} for different parameter settings are presented in Table 3.

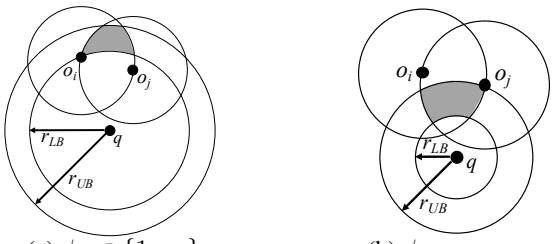
Specifically, we have the following lemma.

Lemma 3. Let o_i and o_j be the object-object distance contributors and o_m be the key query-object distance contributors of the set S to be constructed. For $cost_{unified}$ with different parameter settings, $d(o_m,q)$ can be lower bounded by r_{LB} and upper bounded by r_{UB}, as shown in Table 3. □

Proof. The proof of r_{LB} is shown as follows. When $\phi_1 \in \{1, \infty\}$, $d(o_m,q) > d_f$ because otherwise S is not a feasible set. For $cost_{SumMax}, cost_{MaxMax}$ and $cost_{Sum}$, we do not need to consider an object o if $d(o,q) < \max\{d(o_i,q), d(o_j,q)\}$ because it can not be the key query-object distance contributor of S by definition. Similarly, for $cost_{MaxMax2}$, we do not need to consider object o if $d(o,q) < d(o_i,o_j)$ because it cannot be the key query-object distance contributor of S. When $\phi_1 = -\infty$, we set $r_{LB} = d_f - d(o_i,o_j)$ because otherwise S is not a feasible set.

The proof of r_{UB} is shown as follows. For $cost_{SumMax}$ and $cost_{MaxMax}$, if S contains an object o with $d(o,q) \geq curCost - d(o_i,o_j)$, it is obvious that $cost(S) \geq curCost$. Similarly, for $cost_{MaxMax2}$ ($cost_{Sum}$), if S contains an object o with $d(o,q) \geq curCost$ ($d(o,q) > curCost - d(o_i,q) - d(o_j,q)$), $cost(S) \geq curCost$. For $cost_{MinMax}$ and $cost_{MinMax2}$, we do not need to consider an object o if $d(o,q) \geq min\{d(o_i,q), d(o_j,q)$ because it can not be the key query-object distance contributor of S by definition. Also, in $cost_{MinMax}$, if $d(o,q) \geq curCost - d(o_i,o_j)$, $cost(S) \geq curCost$. □

With the above search strategy introduced, we present the *Unified-E* algorithm in Algorithm 1. Specifically, we maintain an object set $curSet$ for storing the best-known solution found so far, which is initialized to $N(q)$ (line 1), and $curCost$ to be the cost of $curSet$ (line 2). Recall that $N(q)$ is a feasible set. Then, we initialize R_S to be $C(q, r_1)$ (line 3) and find a set P of all object pairs (o_i, o_j) where o_i and o_j are in R_S to take the roles of object-object distance contributors (line 4).



(a) $\phi_1 \in \{1, \infty\}$ (b) $\phi_1 = -\infty$

Fig. 1: Pruning at Step 2 of Unified-E

Algorithm 1 A Unified Approach (An exact algorithm)

Input: A query q, a set $\mathcal{O}$ of objects and a unified cost function $cost_{unified}(S|\alpha, \phi_1, \phi_2)$

1: $curSet \leftarrow N(q)$
2: $curCost \leftarrow cost(curSet)$
3: $R_S \leftarrow C(q, r_1)$
4: $P \leftarrow$ a set of all relevant object pairs (o_i, o_j) where $o_i, o_j \in R_S$ and $d_{LB} \leq d(o_i,o_j) < d_{UB}$
5: **for** each $(o_i, o_j) \in P$ in ascending order of $cost(\{o_i, o_j\})_{LB}$ **do**
6: **if** $cost(\{o_i, o_j\})_{LB} > curCost$ **then**
7: **break;**
8: $R_{ij} \leftarrow C(o_i, d(o_i,o_j)) \cap C(o_j, d(o_i,o_j))$
9: $\mathcal{T} \leftarrow$ a set of all relevant objects $o_m \in R_{ij}$ where $r_{LB} \leq d(o_m,q) \leq r_{UB}$
10: **for** each $o_m \in \mathcal{T}$ in ascending order of $d(o_m,q)$ **do**
11: $S' \leftarrow$ findBestFeasibleSet(o_i, o_j, o_m)
12: **if** $S' \neq \emptyset$ and $cost(S') < curCost$ **then**
13: $curSet \leftarrow S'$
14: $curCost \leftarrow cost(S')$
15: **return** $curSet$

Second, we perform an iterative process as follows. Consider one iteration. We check whether the lower bound of the set containing o_i and o_j is larger than $curCost$ (line 6). If yes, we stop the iterations (line 7). Otherwise, we proceed to initialize the re-

Algorithm 2 findBestFeasibleSet(o_i, o_j, o_m)

Input: Three objects o_i, o_j, o_m
Output: The feasible set (if any) containing o_i, o_j, o_m with the smallest cost

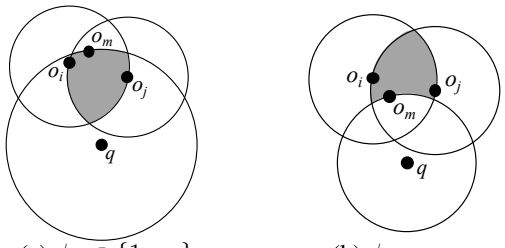
1: $S' \leftarrow \emptyset$
2: $\psi \leftarrow q.\psi - (o_i.\psi \cup o_j.\psi \cup o_m.\psi)$
3: **if** $\psi = \emptyset$ **then**
4: **return** $\{o_i, o_j, o_m\}$
5: **if** $\phi_1 = -\infty$ **then**
6: $R \leftarrow C(o_i, d(o_i, o_j)) \cap C(o_j, d(o_i, o_j)) - C(o_m, d(o_m, q))$
7: **else**
8: $R \leftarrow C(o_i, d(o_i, o_j)) \cap C(o_j, d(o_i, o_j)) \cap C(o_m, d(o_m, q))$
9: $\mathcal{O}' \leftarrow$ a set of all relevant objects in R
10: **if** $\mathcal{O}'$ does not cover ψ **then**
11: **return** $\emptyset$
12: **for** each subset S'' of $\mathcal{O}'$ with $|S''| \leq |\psi|$ **do**
13: **if** S'' covers ψ **then**
14: $S'' \leftarrow S'' \cup \{o_i, o_j, o_m\}$
15: **if** $cost(S'') < cost(S')$ **then**
16: $S' \leftarrow S''$
17: **return** S'

gion R_{ij} to $C(o_i, d(o_i, o_j)) \cap C(o_j, d(o_i, o_j))$ (line 8) and find a set $\mathcal{T}$ of all objects o_m where o_m is in R_{ij} to take the role of key query-object distance contributor (line 9).

Third, we invoke a procedure called *findBestFeasibleSet* (discussed later) for constructing a feasible set S' which takes o_i and o_j as the object-object distance contributors and o_m as the key query-object distance contributor wrt S' (line 11). Then, we update $curSet$ to S' if S' exists and $cost(S') < curCost$ (lines 12 - 14).

Fourth, we iterate the process with the next relevant object in R_{ij} and with the next object pair from R_S until all relevant objects in R_S have been processed.

Next, we introduce the "findBestFeasibleSet" procedure (used in Algorithm 1), which takes three objects o_i, o_j and o_m as input and finds the best feasible set S' (if any) with the *smallest* cost among all feasible sets which have o_i and o_j as the object-object distance contributors have o_m as a key query-object distance contributor. The procedure is presented in Algorithm 2, and it works as follows. First, it initializes S' as an empty set (line 1). Then, it initializes a variable ψ, denoting the set of keywords in $q.\psi$ not covered by S' yet, as $q.\psi - (o_i.\psi \cup o_j.\psi \cup o_m.\psi)$ (line 2). If $\psi = \emptyset$, then it returns $\{o_i, o_j, o_m\}$ immediately (lines 3-4). Otherwise, it proceeds to retrieve the set $\mathcal{O}'$ containing all relevant objects in R, where R is defined based on the value of ϕ_1 (lines 5-9). When $\phi_1 \in \{1, \infty\}$, $R = C(o_i, d(o_i, o_j)) \cap C(o_j, d(o_i, o_j)) \cap C(o_m, d(o_m, q))$ (line 6), and the region is shown in Figure 2(a). When $\phi_1 = -\infty$, $R = C(o_i, d(o_i, o_j)) \cap C(o_j, d(o_i, o_j)) - C(o_m, d(o_m, q))$ (line 8), and the region is shown in Figure 2(b). The major idea of the region R is that including any object outside the region would violate one or both of the following constraints: (1) o_m is the key query-object distance contributor of the set to be found and (2) o_i and o_j are the object-object distance contributors of the set to be found. If $\mathcal{O}'$ does not cover ψ, it returns $\emptyset$ immediately which implies that no such feasible set could be found (lines 10-11). Otherwise, it finds the target by enumerating all possible subsets S'' of $\mathcal{O}'$ with size at most $|\psi|$ (by utilizing the inverted lists maintained for each keyword in ψ), and for each possible S'', if it covers ψ and $cost(S'' \cup \{o_i, o_j, o_m\}) < cost(S')$, S' is updated correspondingly (lines 12-16).



(a) $\phi_1 \in \{1, \infty\}$ (b) $\phi_1 = -\infty$

Fig. 2: Search space R in Algorithm 2

We also develop some other pruning techniques based on a concept of "dominance" for further improving the efficiency of the algorithm. The major idea is that under some parameter settings, the solution of the CoSKQ problem contains only those objects that are not dominated by other objects. Details could be found in Appendix C.

Time complexity analysis. Let $|P|$ be the number of object pairs in P. Note that $|P|$ is usually much smaller than $|\mathcal{O}_q|^2$ since $|P|$ corresponds to the number of relevant objects we process in R_S and the area occupied by R_S is typically small. Let $|R_{ij}|$ be the number of relevant objects in R_{ij}. The time complexity of Algorithm 1 is $O(|P| \cdot |R_{ij}| \cdot \theta)$, where θ is the time complexity of Algorithm 2. It could be verified that θ is dominated by the step of enumerating the object sets (lines 12-16 in Algorithm 2), whose cost is $O(|\mathcal{O}'|^{|q.\psi|-3} \cdot |\psi|^2)$ since it searches at most $O(|\mathcal{O}'|^{|q.\psi|-3})$ subsets S'' that cover ψ and the checking cost for each subset is $O(|\psi|^2)$. As a result, the time complexity of *Unified-E* is $O(|P| \cdot |R_{ij}| \cdot |\mathcal{O}'|^{|q.\psi|-3} \cdot |\psi|^2)$.

4.2 An Approximate Algorithm

In this part, we introduce the approximate algorithm *Unified-A*. Compared with *Unified-E*, *Unified-A* drops the step of object-object distance contributors finding and replaces the step of best feasible set construction which is expensive with a step of (arbitrary) feasible set construction which is efficient, and thus it enjoys significantly better efficiency. Specifically, the *Unified-A* adopts the following search strategy.

- Step 1 (Key Query-Object Distance Contributor Finding): Select a relevant object o to be key query-object distance contributor wrt a set S' to be constructed;
- Step 2 (Feasible Set Construction): Construct the set S' (which has o as a key query-object distance contributor);
- Step 3 (Optimal Set Updating): Update the current best solution $curSet$ if $cost(S') < curCost$, where $curCost$ is the cost of $curSet$;
- Step 4 (Iterative Step): Repeat Step 1 to Step 3 until all possible key query-object distance contributors are iterated.

The above search strategy makes quite effective pruning possible at both Step 1 and Step 2.

Pruning at Step 1. The major idea is that not each relevant object is necessary to be considered as a key query-object distance contributor wrt S' to be constructed. Specifically, in the case of $\phi_1 \in \{1, \infty\}$, all those relevant objects o with $d(o, q) > curCost$ (this is because all those feasible sets S' with o as a key query-object distance contributor have the cost larger than the best-known cost $curCost$, and thus they could be pruned) or $d(o, q) < \max_{o \in N(q) d(o,q)}$ (this is because there exist no feasible sets within the disk of $C(q, \max_{o \in N(q)} d(o, q) - \epsilon)$ where ϵ is close to zero)

Algorithm 3 A Unified Approach (An approximate algorithm)

Input: A query q, a set $\mathcal{O}$ of objects and a unified cost function $cost_{unified}(S|\alpha, \phi_1, \phi_2)$
1: $curSet \leftarrow N(q)$
2: $curCost \leftarrow cost(curSet)$
3: Initialize the region R
4: **for** each relevant object $o \in R$ in ascending order of $d(o, q)$ **do**
5: Initialize the region R_o
6: $S' \leftarrow$ findFeasibleSet(o, R_o)
7: **if** $S' \neq \emptyset$ and $cost(S') < curCost$ **then**
8: $curSet \leftarrow S'$
9: $curCost \leftarrow cost(S')$
10: **return** $curSet$

could be pruned. Therefore, we can maintain a region R which corresponds to the "ring region" enclosed by $C(q, curCost)$ and $C(q, \max_{o \in N(q)} d(o, q))$ for pruning the search space at Step 1. In the case of $\phi_1 = -\infty$, the region R could also be defined correspondingly. Details of the region R for different parameter settings are presented in Table 4.

Pruning at Step 2. We define a region R_o by the key query-object distance contributor found in Step 1 and only the objects in the region need to be considered for constructing S'. The major idea of the pruning is that not all possible objects in R_o are necessary to be considered. Specifically, in the case of $\phi_1 \in \{1, \infty\}$, all those relevant objects outside $C(q, d(o, q))$ could be safely pruned (this is because including one such object would fail o to be a key query-object distance contributor wrt S'). Thus, we can maintain a region R_o which corresponds to $C(q, d(o, q))$ for pruning the search space at Step 2. In the case of $\phi_1 = -\infty$, the region R_o could also be maintained appropriately. Details of the region R_o for different parameter settings are presented in Table 4 as well.

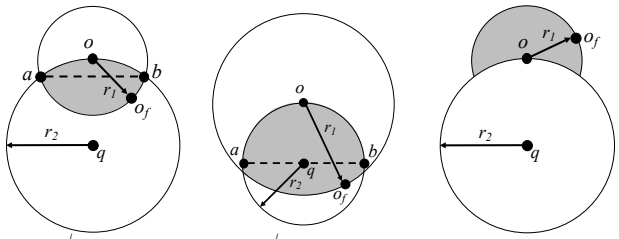
With the above search strategy and pruning techniques introduced, the *Unified-A* algorithm is presented in Algorithm 3. Specifically, we maintain an object set $curSet$ for storing the best-known solution found so far, which is initialized to $N(q)$ (line 1) and $curCost$ to be the cost of $curSet$ (line 2). Then, we perform an iterative process for each relevant object $o \in R$ in ascending order of $d(o, q)$ (lines 3-4). Consider one iteration. First, we initialize the region R_o (line 5). Second, we invoke a procedure called *findFeasibleSet* (discussed later) for constructing a feasible set S' which takes o as a key query-object distance contributor wrt S' (line 6). Third, we update $curSet$ to S' and $curCost$ to $cost(S')$ if S' exists and $cost(S') < curCost$ (lines 7-9). We iterate the process with the next relevant object from R which has not been processed until all relevant objects in R have been processed.

Next, we introduce the "findFeasibleSet" procedure (used in Algorithm 3), which takes an object o and a region R_o as input and finds a feasible set S' (if any) which contains objects in R_o (including o) and has o as a key query-object distance contributor. The procedure is presented in in Algorithm 4, and it is similar to the "findBestFeasibleSet" procedure (in Algorithm 2) except that it replaces the enumeration process with an iterative process (lines 8-14) for searching for a feasible set.

Depending on the value of ϕ_1, the algorithm uses different criterion for picking an object at an iteration, which is described as follows.

Algorithm 4 findFeasibleSet(o, R_o)

Input: An object o, a region R_o
Output: A feasible set (if any) containing objects in R_o (including o)
1: $S' \leftarrow \{o\}$
2: $\psi \leftarrow q.\psi - o.\psi$
3: **if** $\psi = \emptyset$ **then**
4: **return** S'
5: $\mathcal{O}' \leftarrow$ a set of all relevant objects in R_o
6: **if** $\mathcal{O}'$ does not cover ψ **then**
7: **return** $\emptyset$
8: **while** $\psi \neq \emptyset$ **do**
9: **if** $\phi_1 = 1$ **then**
10: $o' \leftarrow \arg\min_{o' \in \mathcal{O}'} \frac{d(o', q)}{|\psi \cap o'.\psi|}$
11: **else**
12: $o' \leftarrow \arg\min_{o' \in \mathcal{O}'} d(o', o)$ and $\psi \cap o'.\psi \neq \emptyset$
13: $S' \leftarrow S' \cup \{o'\}$
14: $\psi \leftarrow \psi - o'.\psi$
15: **return** S'



(a) $\phi_1 = \infty$, $r_1 \leq \sqrt{2} r_2$ (b) $\phi_1 = \infty$, $r_1 > \sqrt{2} r_2$ (c) $\phi_1 = -\infty$

Fig. 3: Illustration of proof of Theorem 2

Case 1: $\phi_1 = 1$. It picks the object which has the smallest ratio of its distance to q to the number of remaining keywords covered. Using this criterion, the algorithm tries to pick objects in a way that minimizes the sum of the distances between the query location and the objects.

Case 2: $\phi_1 \in \{\infty, -\infty\}$. It picks the object which is the nearest to o and covers some of the uncovered keywords. Using this criterion, the algorithm tries to pick objects in a way that minimizes the maximum pairwise distance between the objects.

We also develop two techniques based on the concept of *information re-use* for implementing the *Unified-A* with better efficiency. The details could be found in Appendix D.

Time complexity analysis. Let $|R|$ be the number of relevant objects in R. It could be verified that the complexity of the "findFeasibleSet" (Algorithm 4) is $O(|\psi| \cdot |\mathcal{O}'| \log |\mathcal{O}'|)$ (note that a heap structure with $|\mathcal{O}'|$ elements could be used and there are at most $O(|\psi|)$ operations based on the heap). Therefore, the time complexity of *Unified-A* is $O(|R| \cdot |\psi| \cdot |\mathcal{O}'| \log |\mathcal{O}'|)$.

Approximation ratio analysis. In general, the *Unified-A* algorithm gives different approximation ratios for different parameter settings, which are given in the following theorem.

Theorem 2. The Unified-A algorithm gives approximation ratios as shown in Table 5 for the CoSKQ problem under different parameter settings. □

Proof. Let o be the key query-object distance contributor wrt the optimal solution S_o. Let S be the solution returned by *Unified-A*. In the following, we analyze the approximation ratio of $cost(S)/cost(S_o)$ with different parameter settings.

Cost function	Parameter			R	R_o
	α	ϕ_1	ϕ_2		
$cost_{SumMax}$	0.5	1	1	$C(q, curCost) - C(q, d_f)$	$C(q, d(o, q))$
$cost_{MaxMax}$	0.5	∞	1	$C(q, curCost) - C(q, d_f)$	$C(q, d(o, q))$
$cost_{MaxMax2}$	0.5	∞	∞	$C(q, curCost) - C(q, d_f)$	$C(q, d(o, q))$
$cost_{MinMax}$	0.5	$-\infty$	1	$C(q, curCost)$	$C(q, curCost) \cap C(o, curCost - d(o, q)) - C(q, d(o, q))$
$cost_{MinMax2}$	0.5	$-\infty$	∞	$C(q, 2 \cdot curCost)$	$C(q, 2 \cdot curCost) \cap C(o, curCost - d(o, q)) - C(q, d(o, q))$
$cost_{Sum}$	1	1	-	$C(q, curCost) - C(q, d_f)$	$C(q, d(o, q))$

$d_f = \max_{o \in N(q)} d(o, q)$

TABLE 4: R and R_o in Unified-A

The algorithm iterates each object in the region R, and from the way we initialize R, there must exists an iteration in *Unified-A* such that it processes o and thus it finds the corresponding feasible set S'. Note that o is the key query-object distance contributor of S' because the other objects in S' are located in the region R_o. We have $cost(S) \leq cost(S')$ because *Unified-A* returns the feasible set with the smallest cost. The following proof shows that $cost(S') \leq \gamma cost(S_o)$ which further implies $cost(S) \leq \gamma cost(S_o)$, where γ is the approximation ratio. We consider three cases based on the values of ϕ_1 as follows.

Case 1. $\phi_1 = 1$. In this case, the approach of the algorithm to pick object to form S' is modified from the approximation algorithm of the Weighted Set Cover (WSC) problem, where the keywords in ψ correspond to elements, the objects correspond to sets, and the distances between the objects and query correspond to the set costs. The proof is based on the approximation properties of the WSC problem. Let $w' = \sum_{o' \in S' \setminus \{o\}} d(o', q)$ and $w_o = \sum_{o' \in S_o \setminus \{o\}} d(o', q)$. We have $w' \leq H_{|\psi|} w_o$ where $\psi = q.\psi - o.\psi$, $|\psi| < |q.\psi|$ and H_k is the k^{th} harmonic number. There are three parameter settings (cost functions) adopt this picking object criterion, the proof are shown as follows.

Case 1(a). $\alpha = 1, \phi_1 = 1$ ($cost_{Sum}$).

$$\begin{aligned} \frac{cost(S'|1,1,\cdot)}{cost(S_o|1,1,\cdot)} &= \frac{\sum_{o' \in S'} d(o', q)}{\sum_{o' \in S_o} d(o', q)} \leq \frac{d(o, q) + w'}{d(o, q) + w_o} \\ &\leq \frac{d(o, q) + H_{|\psi|} w_o}{d(o, q) + w_o} \leq H_{|\psi|} \end{aligned}$$

Thus, the approximation ratio is not larger than $H_{|\psi|}$, where $|\psi| < |q.\psi|$, when $cost_{Sum}$ is used.

Case 1(b). $\alpha = 0.5, \phi_1 = 1, \phi_2 = 1$ ($cost_{SumMax}$).

$$\begin{aligned} \frac{cost(S'|0.5,1,1)}{cost(S_o|0.5,1,1)} &= \frac{\sum_{o \in S'} d(o, q) + \max_{o,o' \in S'} d(o, o')}{\sum_{o \in S_o} d(o, q) + \max_{o,o' \in S_o} d(o, o')} \\ &\leq \frac{d(o, q) + w' + d(o_1, q) + d(o_2, q)}{d(o, q) + w_o} \\ &\leq \frac{2(d(o, q) + w')}{d(o, q) + w_o} \\ &\leq \frac{2(d(o, q) + H_{|\psi|} w_o)}{d(o, q) + w_o} \leq 2H_{|\psi|} \end{aligned}$$

where $(o_1, o_2) = \arg\max_{o,o' \in S'} d(o, o')$ and $d(o_1, o_2) \leq d(o_1, q) + d(o_2, q)$ by triangle inequality.

Thus, the approximation ratio is not larger than $2H_{|\psi|}$, where $|\psi| < |q.\psi|$, when $cost_{SumMax}$ is used.

Case 1(c). $\alpha = 0.5, \phi_1 = 1, \phi_2 = \infty$ ($cost_{SumMax2}$).

As proven in Lemma 4, $cost_{SumMax2}$ is equivalent to $cost_{Sum}$. Thus, the approximation bound in this case is same as that of $cost_{Sum}$, which is $H_{|\psi|}$.

Case 2. $\phi_1 = \infty$**.** There are three parameter settings in this case. *Unified-A* can obtain optimal solution when $\alpha = 1$ (i.e. $cost_{Max}$). Next, we discuss the case when $\alpha = 0.5$.

The proof is modified from that of [17]. Let o_f be the object in S' that is farthest from o, $r_1 = d(o_f, o)$ and $r_2 = d(o, q)$. It could be verified that all objects in S' fall in $C(o, r_1) \cap C(q, r_2)$. Besides, it could be verified that $\max_{o \in S_o} d(o, q) = r_2$ and $\max_{o,o' \in S_o} d(o, o') \geq r_1$, where S_o is the optimal solution. Therefore, we know that $cost(S_o|0.5, \infty, \phi_2) \geq (r_1^{\phi_2} + r_2^{\phi_2})^{\frac{1}{\phi_2}}$.

In the following, we consider two cases based on the relationship between r_1 and r_2. It could be verified that $r_1 > \sqrt{2} r_2$ if the diameter of $C(q, r_2)$ falls in $C(o, r_1) \cap C(q, r_2)$. Otherwise we have $r_1 \leq \sqrt{2} r_2$.

Case (i): $r_1 \leq \sqrt{2} r_2$. We denote the intersection points between the boundaries of $C(o, r_1)$ and $C(q, r_2)$ by a and b, as shown in Figure 3(a). It is observed that $\max_{o,o' \in S'} d(o, o') \leq d(a, b)$ because all objects in S' are located in $C(o, r_1) \cap C(q, r_2)$. It could be verified that $d(a, b) = 2\sqrt{r_1^2 - r_1^4/4r_2^2}$. Then, $cost(S') \leq [r_2^{\phi_2} + (2\sqrt{r_1^2 - r_1^4/4r_2^2})^{\phi_2}]^{\frac{1}{\phi_2}}$ .Therefore,

$$\begin{aligned} \frac{cost(S'|0.5, \infty, \phi_2)}{cost(S_o|0.5, \infty, \phi_2)} &\leq \frac{[r_2^{\phi_2} + (2\sqrt{r_1^2 - r_1^4/4r_2^2})^{\phi_2}]^{\frac{1}{\phi_2}}}{r_1^{\phi_2} + r_2^{\phi_2}} \\ &\leq \frac{[\frac{r_2}{r_1}^{\phi_2} + (2\sqrt{1 - r_1^2/4r_2^2})^{\phi_2}]^{\frac{1}{\phi_2}}}{1 + \frac{r_2}{r_1}^{\phi_2}} \end{aligned}$$

Let $z = \frac{r_1}{r_2}$,

$$\begin{aligned} \frac{cost(S'|0.5, \infty, \phi_2)}{cost(S_o|0.5, \infty, \phi_2)} &\leq \frac{[\frac{1}{z^{\phi_2}} + (2\sqrt{1 - z^2/4})^{\phi_2}]^{\frac{1}{\phi_2}}}{1 + \frac{1}{z^{\phi_3}}} \\ &\leq \frac{[1 + (z\sqrt{4 - z^2})^{\phi_2}]^{\frac{1}{\phi_2}}}{1 + z^{\phi_2}} \end{aligned}$$

When $\phi_2 = 1$, we define $f(z) = \frac{1 + z\sqrt{4 - z^2}}{1 + z}$ on $\{z | z \in (0, \sqrt{2}]\}$ because $r_1 \leq \sqrt{2} r_2$. It could be verified that $f(z)$ is monotonically increasing on $(0, 0.875)$ and is monotonically decreasing on $(0.875, \sqrt{2}]$. Thus, $f(z) \leq f(0.875) < 1.375$.

When $\phi_2 = \infty$, we define $g(z) = \frac{\max\{1, z\sqrt{4 - z^2}\}}{\max\{1, z\}}$ on $\{z | z \in (0, \sqrt{2}]\}$. It could be verified that $g(z)$ is monotonically increasing on $(0, 1)$ and is monotonically decreasing on $(1, \sqrt{2}]$. Thus, $g(z) \leq g(1) < \sqrt{3}$.

Case (ii): $r_1 > \sqrt{2} r_2$. Let $diam = 2r_2$ be the diameter of $C(q, r_2)$ and falls in $C(o, r_1) \cap C(q, r_2)$, as shown in Figure 3(b). Similar to case 1, it could be verified that $\max_{o,o' \in S'} d(o, o') \leq diam = 2r_2$. Therefore,

$$\frac{cost(S'|0.5, \infty, \phi_2)}{cost(S_o|0.5, \infty, \phi_2)} \leq \frac{[r_2^{\phi_2} + (2r_2)^{\phi_2}]^{\frac{1}{\phi_2}}}{r_1^{\phi_2} + r_2^{\phi_2}} \leq \frac{[1^{\phi_2} + 2^{\phi_2}]^{\frac{1}{\phi_2}}}{\sqrt{2}^{\phi_2} + 1^{\phi_2}}$$

It could be verified that $\frac{1+2}{\sqrt{2}+1} \leq 1.25$ when $\phi_2 = 1$. When $\phi_2 = \infty$, we have $\frac{\max\{1,2\}}{\max\{\sqrt{2},1\}} \leq \sqrt{2}$.

Based on the above analysis, we can obtain the approximation bounds of the two sub-cases as follows.

Cost function	Parameter			Unified-A	Best known
	α	ϕ_1	ϕ_2	Appro. ratio	Appro. ratio
$cost_{MinMax}$	0.5	$-\infty$	1	2	3 [2]
$cost_{MinMax2}$	0.5	$-\infty$	∞	2	N.A.
$cost_{Sum}$	1	1	-	$H_{\|\psi\|}$	$H_{\|q.\psi\|}$ [2]
$cost_{SumMax}$	0.5	1	1	$2H_{\|q.\psi\|}$	N.A.
$cost_{SumMax2}$	0.5	1	∞	$H_{\|\psi\|}$	$H_{\|q.\psi\|}$ [2]
$cost_{MaxMax}$	0.5	∞	1	1.375	1.375 [17]
$cost_{MaxMax2}$	0.5	∞	∞	$\sqrt{3}$	$\sqrt{3}$ [17]
$cost_{Max}$	1	∞	-	1	N.A.
$cost_{Min}$	1	$-\infty$	-	1	N.A.

TABLE 5: Approx. ratios of Unified-A and existing solutions

Case 2(a). $\alpha = 0.5, \phi_1 = \infty, \phi_2 = 1$ ($cost_{MaxMax}$). The approximation ratio of the algorithm is not larger than $\max\{1.375, 1.25\} = 1.375$.

Case 2(b). $\alpha = 0.5, \phi_1 = \infty, \phi_2 = \infty$ ($cost_{MaxMax2}$). The approximation ratio of the algorithm is not larger than $\max\{\sqrt{3}, \sqrt{2}\} = \sqrt{3}$.

Case 3. $\phi_1 = -\infty$**.** There are three parameter settings in this case. *Unified-A* can obtain optimal solution when $\alpha = 1$ (i.e. $cost_{Min}$). In the following, we discuss the case when $\alpha = 0.5$. Let o_f be the object in S' that is farthest from o, $r_1 = d(o_f, o)$ and $r_2 = d(o, q)$, as shown in Figure 3(c). Besides, it could be verified that $\max_{o,o' \in S_o} d(o, o') \geq r_1$, where S_o is the optimal solution. Thus, we have

$$\frac{cost(S'|0.5, -\infty, \phi_2)}{cost(S_o|0.5, -\infty, \phi_2)} \leq [\frac{r_2^{\phi_2} + (2r_1)^{\phi_2}}{r_2^{\phi_2} + r_1^{\phi_2}}]^{\frac{1}{\phi_2}}$$

The approximation bounds of the two sub-cases are shown as follows.

Case 3(a). $\alpha = 0.5, \phi_1 = -\infty, \phi_2 = 1$ ($cost_{MinMax}$).

$$\frac{cost(S'|0.5, -\infty, 1)}{cost(S_o|0.5, -\infty, 1)} \leq 2 - \frac{r_2}{r_2 + r_1} \leq 2$$

The approximation ratio is not larger than 2 in this case.

Case 3(b). $\alpha = 0.5, \phi_1 = -\infty, \phi_2 = \infty$ ($cost_{MinMax2}$).

We consider the following 3 sub-cases.

Case (i): $r_2 > 2r_1$

$$\frac{cost(S'|0.5, -\infty, \infty)}{cost(S_o|0.5, -\infty, \infty)} \leq \frac{r_2}{r_2} \leq 1$$

Case (ii): $2r_1 \geq r_2 > r_1$

$$\frac{cost(S'|0.5, -\infty, \infty)}{cost(S_o|0.5, -\infty, \infty)} \leq \frac{2r_1}{r_2} \leq 2$$

Case (iii): $r_1 \geq r_2$

$$\frac{cost(S'|0.5, -\infty, \infty)}{cost(S_o|0.5, -\infty, \infty)} \leq \frac{2r_1}{r_1} \leq 2$$

Thus, the approximation ratio is not larger than 2. □

According to the results in Table 5, we know that in despite of the fact that our unified approach is designed for a unified cost function which could be instantiated to many different cost functions, the approximate algorithm based on the unified approach provides *better* (same) approximation ratios than (as) the state-of-the arts for three (two) existing cost functions.

	Hotel	GN	Web
Number of objects	20,790	1,868,821	579,727
Number of unique words	602	222,409	2,899,175
Number of words	80,645	18,374,228	249,132,883

TABLE 6: Datasets used in the experiments

Cost function	Exact Algorithm	Appro. Algorithm
$cost_{MinMax}$	Cao-E1 [2]	Cao-A1 [2]
$cost_{MinMax2}$	Cao-E1 [2]*	Cao-A1 [2]*
$cost_{Sum}$	Cao-E2 [2]	Cao-A3 [2]
$cost_{SumMax}$	Cao-E1 [2]*	Cao-A3 [2]*
$cost_{MaxMax}$	Cao-E1 [2], Long-E [17]	Cao-A1 [2], Cao-A2 [2], Long-A [17]
$cost_{MaxMax2}$	Cao-E1 [2]*, Long-E [17]	Cao-A1 [2]*, Cao-A2 [2]*, Long-A [17]

TABLE 7: Algorithms for comparison (those with the asterisk symbol are adaptations)

5 EMPIRICAL STUDIES

5.1 Experimental Set-up

Datasets. Following the existing studies [3], [17], [2], we used three real datasets in our experiments, namely Hotel, GN and Web. Dataset Hotel contains a set of hotels in the U.S. (www.allstays.com), each of which has a spatial location and a set of words that describe the hotel (e.g., restaurant, pool). Dataset GN was collected from the U.S. Board on Geographic Names (geonames.usgs.gov), where each object has a location and also a set of descriptive keywords (e.g., a geographic name such as valley). Dataset Web was generated by merging two real datasets. One is a spatial dataset called TigerCensusBlock[2], which contains a set of census blocks in Iowa, Kansas, Missouri and Nebraska. The other is WEBSPAM-UK2007[3], which consists of a set of web documents. Table 6 shows the statistics of the three datasets.

Query Generation. Let O be a dataset of objects. Given an integer k, we generate a query q with k query keywords similarly as [3], [17] did. Specifically, to generate $q.\lambda$, we randomly pick a location from the MBR of the objects in O, and to generate $q.\psi$, we first rank all the keywords that are associated with objects in O in descending order of their frequencies and then randomly pick k keywords in the percentile range of [10, 40].

Cost functions. We study all instantiations of our unified cost function except for $cost_{Min}$ and $cost_{SumMax2}$ since as we mentioned in Section 3, the former is of no interest and the latter is equivalent to $cost_{Sum}$. That is, we study 7 cost functions in total, namely $cost_{MinMax}$, $cost_{MinMax2}$, $cost_{Sum}$ and $cost_{SumMax2}$, $cost_{MaxMax}$, $cost_{MaxMax2}$ and $cost_{Max}$.

Algorithms. Both the *Unified-E* algorithm and the *Unified-A* algorithm are studied. For comparison, for the CoSKQ problem with an existing cost function, the state-of-the-art algorithms are used and for the CoSKQ problem with a new cost function, some adaptions of existing algorithms are used. The state-of-the-art algorithms are presented in Table 7, where *Cao-E1*, *Cao-E2*, *Cao-A1*, *Cao-A2* and *Cao-A3* refer to the algorithms *MAXMAX-Exact*, *SUM-Exact*, *MAXMAX-Appro1*, *MAXMAX-Appro2* and *SUM-Appro* [2], respectively, and *Long-E* and *Long-A* refer to the algorithms *MaxSum-Exact* and *MaxSum-Appro* [17], respectively. Note that though the cost function $cost_{SumMax}$ was proposed in [2], it was left as future work to develop solutions and thus we adapt some existing algorithms for the CoSKQ problem with this cost function.

All experiments were conducted on a Linux platform with a 2.66GHz machine and 32GB RAM. The IR-tree index structure is memory resident.

2. http://www.rtreeportal.org
3. http://barcelona.research.yahoo.net/webspam/datasets/uk2007

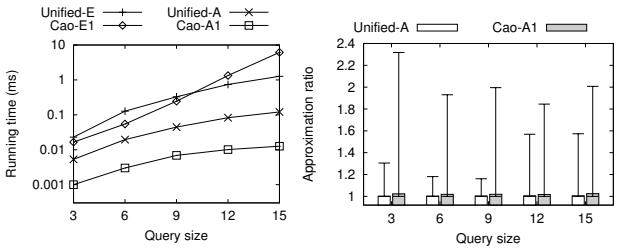
5.2 Experimental Results

Following the existing studies [3], [17], [2], we used the running time and the approximation ratio (for approximate algorithms only) as measurements. Note that different sets of objects with the same costs are treated equally, and thus precision or recall are not used as measures in our experiments. For each experimental setting, we generated 500 queries and ran the algorithms with each of these queries. The average, minimum, and maximum approximation ratios were recorded and shown with bar charts.

5.2.1 Effect of $|q.\psi|$

Following the existing studies [3], [17], we vary the number of query keywords (i.e., $|q.\psi|$) from $\{3, 6, 9, 12, 15\}$. The results on the dataset Hotel are presented and those on the datasets GN and Web are similar and could be found in Appendix E.

(1) $cost_{MinMax}$**.** The results for $cost_{MinMax}$ on the dataset Hotel are shown in Figure 4. According to Figure 4(a), the running time of each algorithm increases when $|q.\psi|$ increases. Our exact algorithm *Unified-E* runs consistently faster than the state-of-the-art algorithm *Cao-E1* and the gap becomes larger when $|q.\psi|$ increases. This could be explained by the fact *Cao-E1* performs the expensive exhaustive search on the pivot objects whose number increases fast with $|q.\psi|$ while *Unified-E* only need to search on the regions that are possible to contain the object sets. Besides, our approximate algorithm *Unified-A* runs quite fast, e.g., less than 0.1 seconds, though it is slower than *Cao-A1*. According to Figure 4(b), *Unified-A* has its approximation ratios consistently better than *Cao-A1*, e.g., the largest approximation ratios of *Unified-A* is *at most* 1.569 while the largest approximation ratios of *Cao-A1* is *at least* 1.845 (and up to 2.317). Note that there could be an significant difference between a solution with 1.569 approximation ratio and that with 2.317 approximation ratio, though it does not seem to look so, e.g., in the case an optimal solution has its cost of 10km, a 1.569-approximate solution has a cost about 16km and a 2.317-approximate solution about 23km, then the difference is about 7km (23km - 16km) which is more than half of the optimal cost. The reason could be that *Unified-A* performs an iterative process on the key query-object distance contributor which helps improve the approximation ratio while *Cao-A1* does not. Besides, we note that the approximation ratio of *Unified-A* is exactly 1 for more than 90% queries, while that of *Cao-A1* is less than 70%.

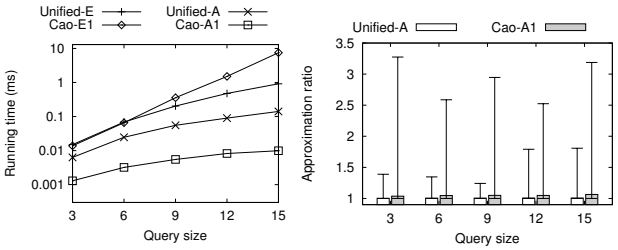


(a) Running time (b) Approximation ratio

Fig. 4: Effect of $|q.\psi|$ on $cost_{MinMax}$ (Hotel)

(2) $cost_{MinMax2}$**.** The results for $cost_{MinMax2}$ on the dataset Hotel are shown in Figure 5, which are similar to those for $cost_{MinMax}$, i.e., *Unified-E* runs consistently faster than *Cao-E1* and *Unified-A* gives better approximation ratios than *Cao-A1* with reasonable efficiency.

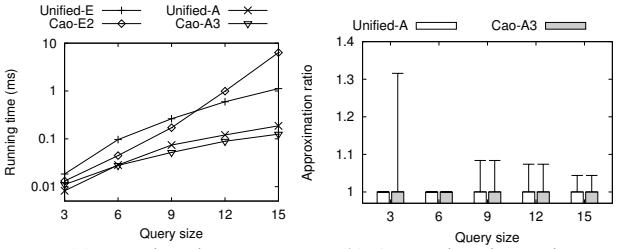
(3) $cost_{Sum}$**.** The results for $cost_{Sum}$ on the dataset Hotel are shown in Figure 6. According to Figure 6(a), *Unified-E* runs similarly fast as *Cao-E2* when $|q.\psi| \leq 9$ and runs faster than *Cao-E2* when $|q.\psi| > 9$. *Unified-E* has a very restrict search space, e.g.,



(a) Running time (b) Approximation ratio

Fig. 5: Effect of $|q.\psi|$ on $cost_{MinMax2}$ (Hotel)

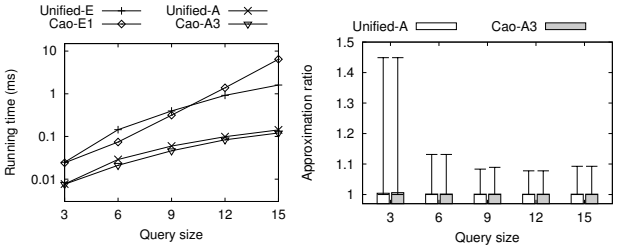
only those dominant objects, and *Cao-E2* is a dynamic programming algorithm which might be more sensitive to $|q.\psi|$. Besides, *Unified-A* has a very similar running time as *Cao-A3*. According to Figure 6(b), *Unified-A* and *Cao-A3* give very similar approximation ratios.



(a) Running time (b) Approximation ratio

Fig. 6: Effect of $|q.\psi|$ on $cost_{Sum}$ (Hotel)

(4) $cost_{SumMax}$**.** The results for $cost_{SumMax}$ on the dataset Hotel are shown in Figure 7, which are similar to those for $cost_{Sum}$ except that the competitor is *Cao-E1*, i.e., *Unified-E* runs faster than *Cao-E1* when $|q.\psi|$ grows and *Unified-A* has similar running time and also approximation ratios as *Cao-A3*.

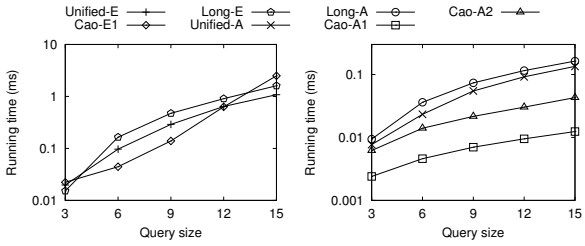


(a) Running time (b) Approximation ratio

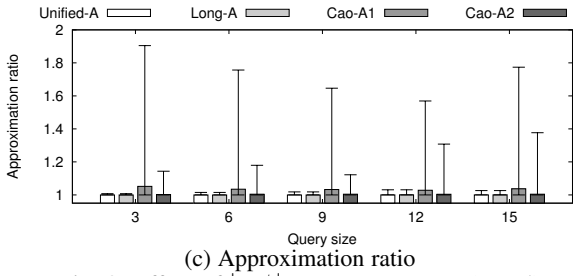
Fig. 7: Effect of $|q.\psi|$ on $cost_{SumMax}$ (Hotel)

(5) $cost_{MaxMax}$**.** The results for $cost_{MaxMax}$ on the dataset Hotel are shown in Figure 8. According to Figure 8(a), each algorithm has its running time grows when $|q.\psi|$ increases (in particular, *Cao-E1* has its running time grows the fastest). Besides, *Unified-E* runs consistently faster than *Long-E* and runs faster than *Cao-E1* as well when $|q.\psi|$ gets larger. According to Figure 8(b), all approximate algorithms including *Unified-A* run fast, e.g., less then 0.1 seconds, and according to Figure 8(c), *Unified-A* is one of two algorithms that give the best approximation ratio (the other is *Long-A*). Note that *Unified-A* runs consistently faster than *Long-A*, and the reason could be that *Unified-A* has computation strategies based on information re-use while *Long-A* does not. The largest approximation ratios of *Unified-A* is only 1.031, while that of *Cao-A1* and *Cao-A2* could be up to 1.904 and 1.377, respectively. Besides, *Unified-A* gives approximation ratio of exactly 1 for 98% queries, while that of *Cao-A1* and *Cao-A2* are 51% and 83%, respectively.

(6) $cost_{MaxMax2}$**.** The results for $cost_{MaxMax2}$ on the dataset Hotel are shown in Figure 9, which are similar as those for $cost_{MaxMax}$, i.e., *Unified-E* has the best efficiency in general



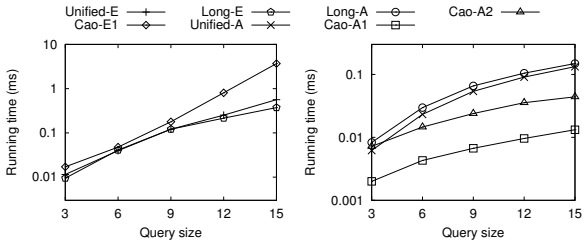
(a) Running time (Exact) (b) Running time (Appro.)



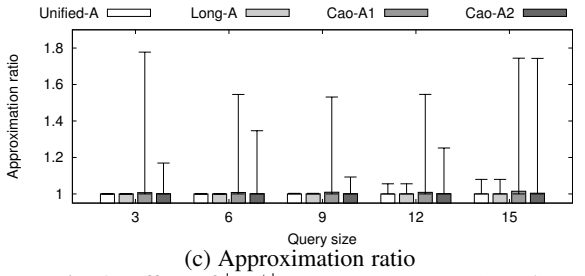
(c) Approximation ratio

Fig. 8: Effect of $|q.\psi|$ on $cost_{MaxMax}$ (Hotel)

and *Unified-A* is among one of the two algorithms which give the best approximation ratios and also run reasonably fast. Note that the largest approximation ratios of *Unified-A* is only 1.080, while that of *Cao-A1* and *Cao-A2* could be up to 1.778 and 1.347, respectively.



(a) Running time (Exact) (b) Running time (Appro.)



(c) Approximation ratio

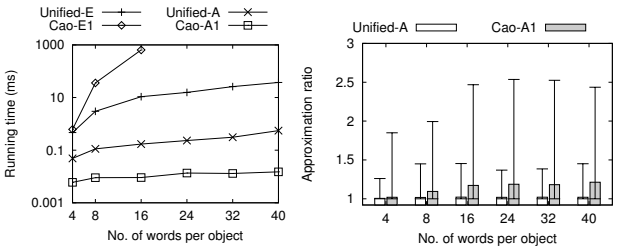
Fig. 9: Effect of $|q.\psi|$ on $cost_{MaxMax2}$ (Hotel)

(7) $cost_{Max}$**.** The results for $cost_{Max}$ are shown in Figure 16(a). According to the results, both *Unified-E* and *Unified-A* run very fast, e.g., they ran less than 0.01 ms for all settings of $|q.\psi|$. This is mainly because that both algorithms essentially find $N(q)$ as the solution.

5.2.2 Effect of average $|o.\psi|$

We further generated 5 datasets based on the Hotel dataset, where the average number of keywords an object contains (i.e. average $|o.\psi|$) is close to 8, 16, 24, 32, and 40, respectively. In the Hotel dataset, the average number of keywords an object contains is close to 4. To generate a dataset with its average $|o.\psi|$ equal to 8, we do the following. For each object o in the Hotel dataset, we augment $o.\psi$ by including all those keywords in $o'.\psi$ to $o.\psi$ (i.e., $o.\psi \leftarrow o.\psi \cup o'.\psi$) where o' is a randomly picked object. To generate the datasets with the average $|o.\psi|$ equal to 16, 24, 32 and 40, we repeat the above process appropriate times. We vary average $|o.\psi|$ from $\{4, 8, 16, 24, 32, 40\}$ and following [2], we use the default setting of $|q.\psi| = 10$.

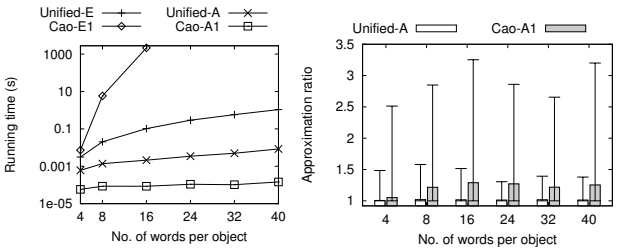
(1) $cost_{MinMax}$**.** The results for $cost_{MinMax}$ are shown in Figure 10, where the results of running time of *Cao-E1* for $|o.\psi| \geq 24$ are not shown simply because it ran for more than 10 hours (this applies for all the following results). According to Figure 10(a), all algorithms except for *Cao-E1* are quite scalable when $|o.\psi|$ grows. The poor scalability of *Cao-E1* could be due to the fact that *Cao-E1* is based on the search space of relevant objects around the candidate objects, which grows rapidly when $|o.\psi|$ increases. Besides, our exact algorithm *Unified-E* runs consistently better than *Cao-E1* and *Unified-A* runs fast, though not as fast as *Cao-A1*, and gives obviously better approximation ratios than *Cao-A1* (Figure 10(b)). Specifically, the largest approximation ratios of *Unified-A* is only 1.454, which is small, while that of *Cao-A1* is up to 2.536, which is not suitable for practical use.



(a) Running time (b) Approximation ratio

Fig. 10: Effect of average $|o.\psi|$ on $cost_{MinMax}$

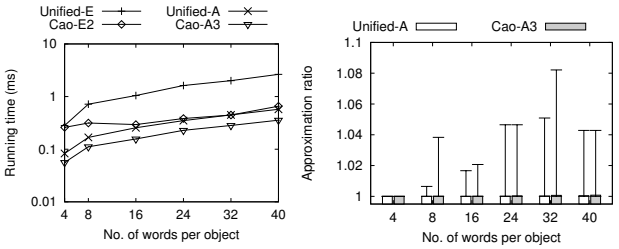
(2) $cost_{MinMax2}$**.** The results for $cost_{MinMax2}$ are shown in Figure 11, which are similar to those for $cost_{MinMax}$, i.e., all algorithms except for *Cao-E1* are scalable when $|o.\psi|$ grows, *Unified-E* runs consistently faster than *Cao-E1*, and *Unified-A* runs fast and gives the best approximation ratios.



(a) Running time (b) Approximation ratio

Fig. 11: Effect of average $|o.\psi|$ on $cost_{MinMax2}$

(3) $cost_{Sum}$**.** The results for $cost_{Sum}$ are shown in Figure 12. According to the Figure 12(a), *Unified-E* runs slower than *Cao-E2*, and the reason is perhaps that the pruning technique of *Unified-E* based on dominant objects becomes less effective when $|o.\psi|$ increases. Besides, *Unified-A* runs slightly slower than *Cao-A3* but gives a better approximation than *Cao-A3* (Figure 12(b)). This is because *Unified-A* construct a feasible set for each key query-object distance contributor and pick the best one as the solution.

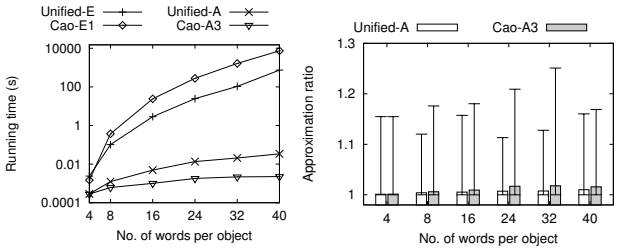


(a) Running time (b) Approximation ratio

Fig. 12: Effect of average $|o.\psi|$ on $cost_{Sum}$

(4) $cost_{SumMax}$**.** Under the default setting of $|q.\psi| = 10$, the running times of all exact algorithms including *Unified-E* and *Cao-E1* grow very rapidly when $|o.\psi|$ increases, e.g., the algorithms

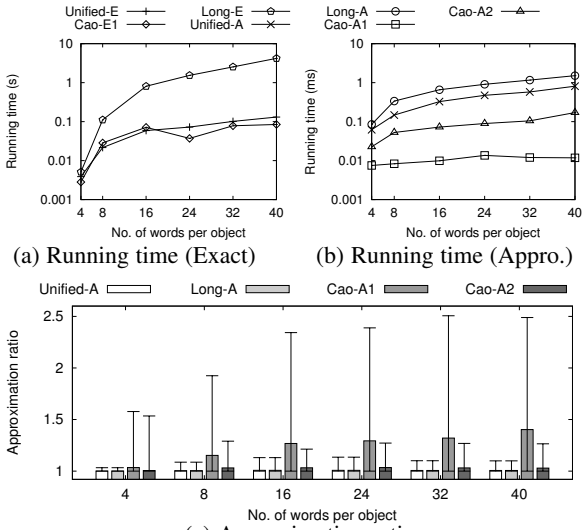
ran for more than 1 day when $|o.\psi| \geq 8$. Thus, for better comparison among the algorithms, we particularly use the setting of $|q.\psi| = 8$ for $cost_{SumMax}$. According to Figure 13(a), *Unified-E* runs consistently faster than *Cao-E1* and *Unified-A* runs fast, though not as fast as *Cao-A3*, and gives a better approximation ratio (Figure 13(b)). Specifically, the largest approximation ratios of *Unified-A* and *Cao-A3* are 1.160 and 1.251, respectively.



(a) Running time (b) Approximation ratio

Fig. 13: Effect of average $|o.\psi|$ on $cost_{SumMax}$

(5) $cost_{MaxMax}$**.** The results for $cost_{MaxMax}$ are shown in Figure 14. According to Figure 14(a), *Unified-E* is one of the two algorithms that run the fastest and the other is *Cao-E1*. According to Figure 14(b) and (c), all approximate algorithms including *Unified-A* run reasonably fast and *Unified-A* is one of the two algorithms which give the best approximation ratios (the other is *Long-A*). Specifically, the largest approximation ratios of *Unified-A* is only 1.135, while that of *Cao-A1* and *Cao-A2* are 2.506 and 1.534, respectively, which are much larger.



(a) Running time (Exact) (b) Running time (Appro.)

(c) Approximation ratio

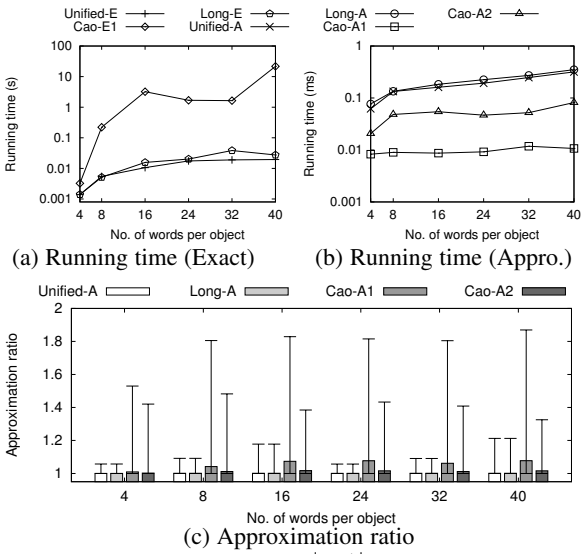
Fig. 14: Effect of average $|o.\psi|$ on $cost_{MaxMax}$

(6) $cost_{MaxMax2}$**.** The results for $cost_{MaxMax2}$ are shown in Figure 15, which are similar to those for $cost_{MaxMax}$, i.e., *Unified-E* is one of the two fastest exact algorithm and *Unified-A* runs reasonably fast and is one of the two algorithms which give the best approximation ratios.

(7) $cost_{Max}$**.** The results for $cost_{Max}$ are shown in Figure 16(b). According to the results, both *Unified-E* and *Unified-A* run very fast, e.g., they ran less than 0.02 ms on all settings of $|o.\psi|$.

5.2.3 Scalability Test

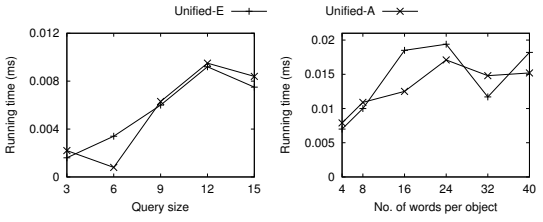
Following the existing studies [3], [17], [2], we generated 5 synthetic datasets for the experiments of scalability test, in which the numbers of objects used are 2M, 4M, 6M, 8M and 10M. Specifically, we generated a synthetic dataset by augmenting the GN



(a) Running time (Exact) (b) Running time (Appro.)

(c) Approximation ratio

Fig. 15: Effect of average $|o.\psi|$ on $cost_{MaxMax2}$

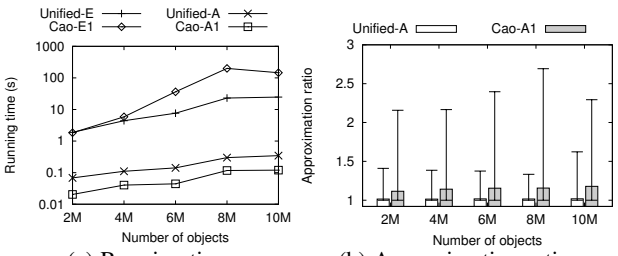


(a) Effect of $|q.\psi|$ (b) Effect of average $|o.\psi|$

Fig. 16: Experiments on $cost_{Max}$

datasets with additional objects as follows. Each time, we create a new object o with $o.\lambda$ set to be a random location from the original GN dataset by following the distribution and $o.\psi$ set to be a random document from GN and then add it into the GN dataset. We vary the number of objects from $\{2M, 4M, 6M, 8M, 10M\}$, following [2], we use the default setting of $|q.\psi| = 10$.

(1) $cost_{MinMax}$**.** The results for $cost_{MinMax}$ are shown in Figure 17. According to Figure 17(a), our exact algorithm *Unified-E* runs consistently faster than *Cao-E1* and it is scalable wrt the number of objects, e.g., it ran within 30 seconds on a dataset with 10M objects. Besides, our approximate algorithm *Unified-A* is also scalable, e.g., it ran within 1 second on a dataset with 10M objects, and gives near-to-optimal approximation ratios (Figure 17(b)). The largest approximation ratios of *Unified-A* is only 1.622, which is very small, while that of *Cao-A1* is 2.692, which is not practical. This also conform with our theoretical analysis that *Unified-A* has a better approximation ratio than *Cao-A1* in $cost_{MinMax}$.



(a) Running time (b) Approximation ratio

Fig. 17: Scalability test on $cost_{MinMax}$

The results for the remaining cost functions are put in Appendix F due to the page limit. According to the results, we know that both *Unified-E* and *Unified-A* are scalable to large datasets.

5.3 Summary Of Experimental Results

Our exact algorithm *Unified-E* is clearly the best exact algorithm for CoSKQ queries not only because it is a *unified* approach but also it is always among those with the best running times (e.g., it beats the state-of-the arts *consistently* for $cost_{MinMax}$ and $cost_{MinMax2}$, *when* $|q.\psi|$ *becomes large* for $cost_{Sum}$ and $cost_{SumMax}$, and *under the majority of settings* for $cost_{MaxMax}$ and $cost_{MaxMax2}$).

Our approximate algorithm *Unified-A* runs reasonably fast (e.g., for the majority settings of $|q.\psi|$, it ran within 0.1 seconds), while sometimes it is not as fast as the competitors because *Unified-A* has some more checking so that it can take care all cost functions. Meanwhile, *Unified-A* is always among the those which give the best approximation ratios close to 1 and runs always faster than those algorithms which give similar approximation ratios as *Unified-A*.

6 Conclusion

In this paper, we proposed a unified cost function for CoSKQ. This cost function expresses all existing cost functions in the literature and a few cost functions that have not been studied before. We designed a unified approach, which consists of one exact algorithm and one approximate algorithm. The exact algorithm runs comparably fast as the existing exact algorithms, while the approximate algorithm provides a comparable approximation ratio as the existing approximate algorithms. Extensive experiments were conducted which verified our theoretical findings.

There are several interesting future research directions. One direction is to design a cost function such that it penalizes those objects with too much keywords for fairness. Another direction is to extend CoSKQ with the unified cost function to other distance metrics such as road networks. It is also interesting to extend the unified approach to handle the route-oriented spatial keyword queries. Besides, it is left as a remaining issue to study the CoSKQ problem with a moving query point.

Acknowledgments

We are grateful to the anonymous reviewers for their constructive comments on this paper. The research of Harry Kai-Ho Chan and Raymond Chi-Wing Wong is supported by HKRGC GRF 16219816.

References

[1] X. Cao, L. Chen, G. Cong, and X. Xiao. Keyword-aware optimal route search. *PVLDB*, 5(11):1136–1147, 2012.
[2] X. Cao, G. Cong, T. Guo, C. S. Jensen, and B. C. Ooi. Efficient processing of spatial group keyword queries. *TODS*, 40(2):13, 2015.
[3] X. Cao, G. Cong, C. S. Jensen, and B. C. Ooi. Collective spatial keyword querying. In *SIGMOD*, pages 373–384. ACM, 2011.
[4] X. Cao, G. Cong, C. S. Jensen, and M. L. Yiu. Retrieving regions of intersect for user exploration. *PVLDB*, 7(9), 2014.
[5] A. Cary, O. Wolfson, and N. Rishe. Efficient and scalable method for processing top-k spatial boolean queries. In *SSDBM*, pages 87–95. Springer, 2010.
[6] H. K.-H. Chan, C. Long, and R. C.-W. Wong. Inherent-cost aware collective spatial keyword queries. In *SSTD*, 2017.
[7] D.-W. Choi, J. Pei, and X. Lin. Finding the minimum spatial keyword cover. In *ICDE*, pages 685–696. IEEE, 2016.
[8] G. Cong, C. S. Jensen, and D. Wu. Efficient retrieval of the top-k most relevant spatial web objects. *PVLDB*, 2(1):337–348, 2009.
[9] G. Cong, H. Lu, B. C. Ooi, D. Zhang, and M. Zhang. Efficient spatial keyword search in trajectory databases. *Arxiv preprint arXiv:1205.2880*, 2012.
[10] K. Deng, X. Li, J. Lu, and X. Zhou. Best keyword cover search. *TKDE*, 27(1):61–73, 2015.
[11] J. Fan, G. Li, L. Z. S. Chen, and J. Hu. Seal: Spatio-textual similarity search. *PVLDB*, 5(9):824–835, 2012.
[12] I. D. Felipe, V. Hristidis, and N. Rishe. Keyword search on spatial databases. In *ICDE*, pages 656–665. IEEE, 2008.
[13] Y. Gao, J. Zhao, B. Zheng, and G. Chen. Efficient collective spatial keyword query processing on road networks. *ITS*, 17(2):469–480, 2016.
[14] T. Guo, X. Cao, and G. Cong. Efficient algorithms for answering the m-closest keywords query. In *SIGMOD*. ACM, 2015.
[15] Z. Li, K. Lee, B. Zheng, W. Lee, D. Lee, and X. Wang. Ir-tree: An efficient index for geographic document search. *TKDE*, 23(4):585–599, 2011.
[16] J. Liu, K. Deng, H. Sun, Y. Ge, X. Zhou, and C. Jensen. Clue-based spatio-textual query. *PVLDB*, 10(5):529–540, 2017.
[17] C. Long, R. C.-W. Wong, K. Wang, and A. W.-C. Fu. Collective spatial keyword queries:a distance owner-driven approach. In *SIGMOD*, pages 689–700. ACM, 2013.
[18] J. Rocha, O. Gkorgkas, S. Jonassen, and K. Nørvåg. Efficient processing of top-k spatial keyword queries. In *SSTD*, pages 205–222. Springer, 2011.
[19] J. B. Rocha-Junior and K. Nørvåg. Top-k spatial keyword queries on road networks. EDBT, pages 168–179. ACM, 2012.
[20] S. Shang, R. Ding, B. Yuan, K. Xie, K. Zheng, and P. Kalnis. User oriented trajectory search for trip recommendation. In *EDBT*, pages 156–167. ACM, 2012.
[21] A. Skovsgaard and C. S. Jensen. Finding top-k relevant groups of spatial web objects. *VLDBJ*, 24(4):537–555, 2015.
[22] S. Su, S. Zhao, X. Cheng, R. Bi, X. Cao, and J. Wang. Group-based collective keyword querying in road networks. *Information Processing Letters*, 118:83–90, 2017.
[23] D. Wu, G. Cong, and C. Jensen. A framework for efficient spatial web object retrieval. *VLDBJ*, 21(6):797–822, 2012.
[24] D. Wu, M. Yiu, G. Cong, and C. Jensen. Joint top-k spatial keyword query processing. *TKDE*, 24(10):1889–1903, 2012.
[25] D. Wu, M. L. Yiu, C. S. Jensen, and G. Cong. Efficient continuously moving top-k spatial keyword query processing. In *ICDE*, pages 541–552. IEEE, 2011.
[26] Y. Zeng, X. Chen, X. Cao, S. Qin, M. Cavazza, and Y. Xiang. Optimal route search with the coverage of users' preferences. In *IJCAI*, pages 2118–2124, 2015.
[27] C. Zhang, Y. Zhang, W. Zhang, and X. Lin. Inverted linear quadtree: Efficient top k spatial keyword search. In *ICDE*, pages 901–912. IEEE, 2013.
[28] D. Zhang, Y. M. Chee, A. Mondal, A. Tung, and M. Kitsuregawa. Keyword search in spatial databases: Towards searching by document. In *ICDE*, pages 688–699. IEEE, 2009.
[29] D. Zhang, B. C. Ooi, and A. K. H. Tung. Locating mapped resources in web 2.0. In *ICDE*, pages 521–532. IEEE, 2010.
[30] D. Zhang, K.-L. Tan, and A. K. H. Tung. Scalable top-k spatial keyword search. EDBT/ICDT, pages 359–370. ACM, 2013.

Harry Kai-Ho Chan received the BEng and MPhil degrees in computer science and engineering from the Hong Kong University of Science and Technology (HKUST) in 2013 and 2015, respectively. He is currently working toward the PhD degree from the Department of Computer Science and Engineering, HKUST. His research interests include database and data mining.

Cheng Long received the PhD degree from the Hong Kong University of Science and Technology (HKUST) in 2015. He is an academic lecturer at Queen's University Belfast (UK). His research interest include database and data mining.

Raymond Chi-Wing Wong received the BSc, MPhil and PhD degrees in computer science and engineering from the Chinese University of Hong Kong (CUHK) in 2002, 2004, and 2008, respectively. He is an associate professsor of the Department of Computer Science and Engineering, the Hong Kong University of Science and Technology. His research interests include database and data mining.

APPENDIX A
EQUIVALENCE OF $cost_{SumMax2}$ AND $cost_{Sum}$

We have the following lemma to show the functionality of $cost_{SumMax2}$ and $cost_{Sum}$ are equivalent.

Lemma 4. Let S be an object set. $cost_{SumMax2}(S) = cost_{Sum}(S)$. □

Proof. Let $(o_1, o_2) = \arg\max_{o_1,o_2 \in S} d(o_1, o_2)$. We have $cost_{SumMax2}(S) = \max\{\sum_{o \in S} d(o, q), d(o_1, o_2)\}$. Note that $d(o_1, q) + d(o_2, q) \geq d(o_1, o_2)$ by triangle inequality and $\sum_{o \in S} d(o, q) \geq d(o_1, q) + d(o_2, q)$. Thus, $cost_{SumMax2}(S) = \sum_{o \in S} d(o, q) = cost_{Sum}(S)$. □

This lemma suggests that it is sufficient to consider one of these two cost functions. In this paper, we focus the discussion on $cost_{Sum}$.

APPENDIX B
PROOF OF THEOREM 1

We first give the decision problem of CoSKQ. Given a set O of spatial objects each $o \in O$ associated with a location $o.\lambda$ and a set of keywords $o.\psi$, a query q consisting of a query location $q.\lambda$ and a set of query keywords $q.\psi$, and a real number C, the problem is to determine whether there exists a set S of objects in O such that S covers the query keywords and $cost_{unified}(S)$ is at most C.

We then prove by transforming the *3-satisfiability* (3-SAT) problem which is known to be NP-Complete to the CoSKQ problem and showing the equivalence between two problems. The description of the 3-SAT problem is given as follows. Let U be a set of literals $\{e_1, \overline{e_1}, ..., e_n, \overline{e_n}\}$ where $\overline{e_i}$ is the negation of e_i. Given an expression $E = C_1 \wedge C_2 \wedge ... \wedge C_m$ where $C_j = x_j \vee y_j \vee z_j$ and $x_j, y_j, z_j \in U$ for $1 \leq j \leq m$, the problem is to determine whether there exists a truth assignment for e_i for $i \leq i \leq n$ such that E is true.

Based on the value of parameter ϕ_1, we use different transformations.

Case 1. $\phi_1 = 1$. We construct a set O of $2n$ objects as follows. For each literal e_i in U, we create an object o_i in O, and for each literal $\overline{e_i}$ in U, we create an object o'_i in O. In total, there are $2n$ objects in O. We set the locations of the objects in O such that they are all located at the same place i.e., for any $o \in O$, $o.\lambda$ is identical. Besides, for each object o_i $(1 \leq i \leq n)$, we set $o_i.\psi$ such that $o_i.\psi$ includes a keyword k_i corresponding to e_i and a keyword k'_j corresponding to C_j if C_j involves e_i for $1 \leq j \leq m$. Similarly, for each object o'_i $(1 \leq i \leq n)$, we set $o'_i.\psi$ such that $o'_i.\psi$ includes the keyword k'_i and all k'_j's with C_j involving $\overline{e_i}$ for $1 \leq j \leq m$. We construct a query q by setting $q.\lambda$ to be a location such that $d(o, q) = 1$ for any object $o \in O$ and $q.\psi$ to be a set of $m + n$ keywords, $\{k_1, k_2, ..., k_n, k'_1, k'_2, ..., k'_m\}$. We set C to be n. The above transformation process could be done in polynomial time.

Case 2. $\phi_1 \in \{\infty, -\infty\}$. We construct a set O of $2n$ objects as follows. For each literal e_i in U, we create an object o_i in O, and for each literal $\overline{e_i}$ in U, we create an object o'_i in O. In total, there are $2n$ objects in O. For the locations of the objects, consider a circle Cir with its center at $q.\lambda$ (which is selected arbitrarily) and its radius equal to 1. We set the locations of the objects in O such that they are all located on the boundary of Cir such that $d(o_i, o'_i) = 2$. Besides, for each object o_i $(1 \leq i \leq n)$, we set $o_i.\psi$ such that $o_i.\psi$ includes a keyword k_i corresponding to e_i and a keyword k'_j corresponding to C_j if C_j involves e_i for $1 \leq j \leq m$. Similarly, for each object o'_i $(1 \leq i \leq n)$, we set $o'_i.\psi$ such that $o'_i.\psi$ includes the keyword k'_i and all k'_j's with C_j involving $\overline{e_i}$ for $1 \leq j \leq m$. We construct a query q by setting $q.\lambda$ arbitrarily and $q.\psi$ to be a set of $m + n$ keywords, $\{k_1, k_2, ..., k_n, k'_1, k'_2, ..., k'_m\}$. The above transformation process could be done in polynomial time. We consider the following subcases for setting C.

Case 2(a). $\phi_2 = 1$. We set $C = 3 - \epsilon$ where ϵ is close to zero.

Case 2(b). $\phi_2 = \infty$. We set $C = 2 - \epsilon$ where ϵ is close to zero.

We show the equivalence between two problem instances as follows. Suppose that the answer of the 3-SAT problem is "yes", i.e., there exists a truth assignment for the literals in U such that E is correct. We denote the truth assignment by a set T of literals which are true under the assignment. Note that T has exactly n literals and e_i and $\overline{e_i}$ do not appear in T simultaneously for any $1 \leq i \leq n$. Then, it could be verified that the set of objects each corresponding to a literal in T covers $q.\psi$ and the cost of the set at most C, and thus the answer of the CoSKQ problem is also "yes". Suppose that the answer of the CoSKQ problem is "yes". Let S be the set of objects in O that covers $q.\psi$ and has the cost at most C. We know that object o_i and o'_i are not included in S simultaneously. It could be verified that with the truth assignment represented by the set of literals corresponding to the objects in S, E is correct, and thus the answer of the 3-SAT problem is also "yes". □

APPENDIX C
PRUNING BASED ON DOMINANCE

To improve the efficiency of the algorithm, we propose a pruning strategy to prune the search space when $\alpha = 1$ and $\phi_1 = 1$. Before we give the strategy, we first introduce the concept of **dominance**. Given a query q, two objects o_1 and o_2, we say o_1 dominate o_2 if the following two conditions are satisfied. (1) $d(o_1, q) < d(o_2, q)$, and (2) all keywords in $q.\psi$ that are covered by o_2 can be covered by o_1, (i.e. $q.\psi \cap o_1.\psi \supseteq q.\psi \cap o_2.\psi$). A dominant object is defined to be an object that is not dominated by any other objects. Then we have the following lemma to prune the objects that are not dominant objects.

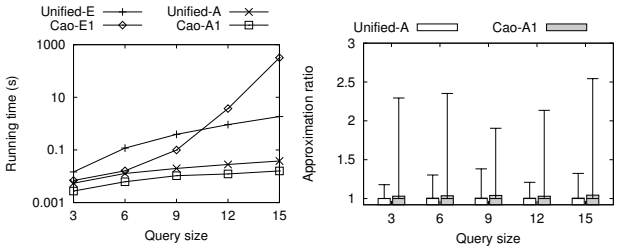
Lemma 5. When $\alpha = 1$ and $\phi_1 = 1$, all objects in the optimal solution S are dominant objects. □

Proof: We prove this by contradiction. Let an object $o \in S$ that is not a dominant object. Then, there must exist an object o' that dominate o. Note that o' also covers the query keywords covered by o and is closer to q. We can construct a better solution $S' = S \setminus \{o\} \cup \{o'\}$, which contradicts the fact that S is the optimal solution. □

Based on this lemma, it is sufficient for the algorithm to consider the dominant objects only when enumerating the object sets. Specifically, whenever the algorithm performs a range query, it discards the objects that are being dominated and proceeds with the dominant objects.

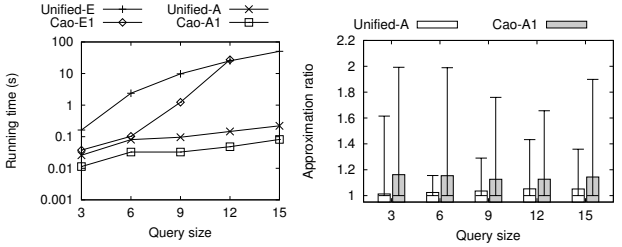
APPENDIX D
BETTER IMPLEMENTATION BASED ON INFORMATION RE-USE

To implement the *Unified-A* algorithm efficiently, we have the following computation strategies. First, when the algorithm finding



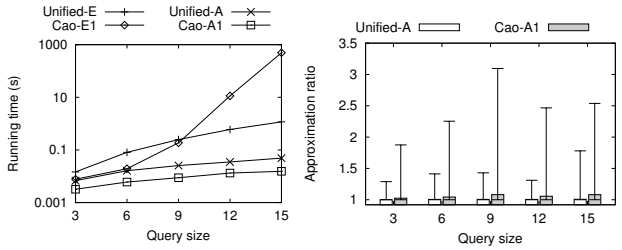
(a) Running time (b) Approximation ratio

Fig. 18: Effect of $|q.\psi|$ on $cost_{MinMax}$ (GN)



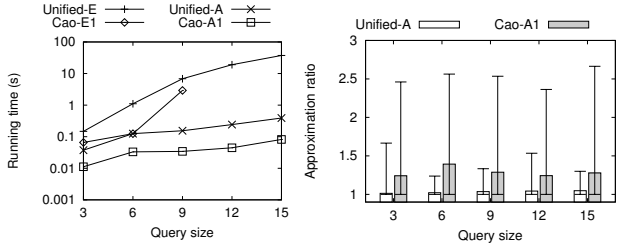
(a) Running time (b) Approximation ratio

Fig. 19: Effect of $|q.\psi|$ on $cost_{MinMax}$ (Web)



(a) Running time (b) Approximation ratio

Fig. 20: Effect of $|q.\psi|$ on $cost_{MinMax2}$ (GN)



(a) Running time (b) Approximation ratio

Fig. 21: Effect of $|q.\psi|$ on $cost_{MinMax2}$ (Web)

the set of all relevant objects in R_o (line 5 in Algorithm 4), instead of issuing a range query in each iteration, it re-uses the information from the previous iteration by maintaining the region R_o dynamically. Specifically, consider one iteration. The algorithm finds a feasible set that has an object o as a key query-object distance contributor in the region R_o. After it finishes the current iteration, it adds o into R_o (when $\phi_1 \in \{1, \infty\}$), or removes o from R_o (when $\phi_1 = -\infty$).

Second, when the algorithm performs the iterative process (lines 8-14 in Algorithm 4), instead of searching for the object with minimum ratio (distance) from $\mathcal{O}'$ in each iteration, it maintains a heap structure for storing the objects. Specifically, when $\phi_1 = 1$, the key of the objects in the heap are the ratios, and the heap is updated after each object is picked. When $\phi_1 \in \{\infty, -\infty\}$, the key of the objects in the heap are the distances, and in each iteration the algorithm picks the relevant object with the smallest distance.

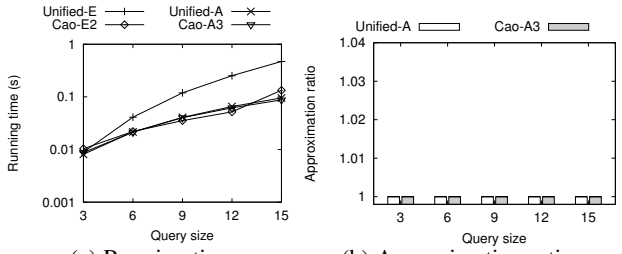
APPENDIX E
EXPERIMENTAL RESULTS ON THE DATASETS GN AND WEB

In the following, we present the experimental results on the datasets GN and Web of varying $|q.\psi|$. Following the existing studies [3], [17], we vary the number of query keywords (i.e., $|q.\psi|$) from $\{3, 6, 9, 12, 15\}$.

(1) $cost_{MinMax}$**.** The results for $cost_{MinMax}$ on the datasets GN and Web are shown in Figure 18 and Figure 19, respectively, which are similar to that on the dataset Hotel. The result of running time of *Cao-E1* for $|q.\psi| = 15$ is not shown in Figure 19 simply because it ran for more than 10 hours (this applies for all the following results).

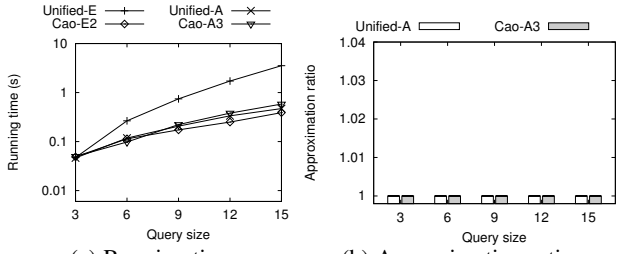
(2) $cost_{MinMax2}$**.** The results for $cost_{MinMax2}$ on the datasets GN and Web are shown in Figure 20 and Figure 21, respectively, which are similar to those for $cost_{MinMax}$.

(3) $cost_{Sum}$**.** The results for $cost_{Sum}$ on the datasets GN and Web are shown in Figure 22 and Figure 23, respectively. According to the results, *Unified-E* runs slower than *Cao-E2* but still within a reasonable time (e.g. within 10 seconds on the largest dataset Web). Besides, *Unified-A* has a very similar running time as *Cao-A3*, while *Unified-A* can always obtain an approximation ratios of 1.



(a) Running time (b) Approximation ratio

Fig. 22: Effect of $|q.\psi|$ on $cost_{Sum}$ (GN)



(a) Running time (b) Approximation ratio

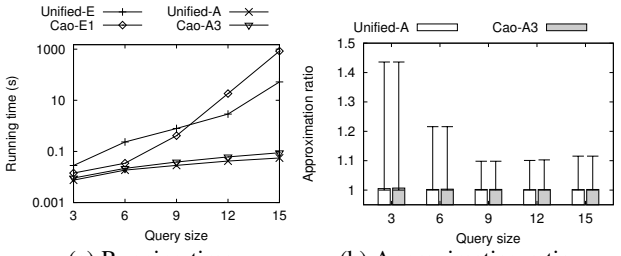
Fig. 23: Effect of $|q.\psi|$ on $cost_{Sum}$ (Web)

(4) $cost_{SumMax}$**.** The results for $cost_{SumMax}$ on the datasets GN and Web are shown in Figure 24 and Figure 25, respectively, which are similar to that on the dataset Hotel.

(5) $cost_{MaxMax}$**.** The results for $cost_{MaxMax}$ on the datasets GN and Web are shown in Figure 26 and Figure 27, respectively, which are similar to that on the dataset Hotel.

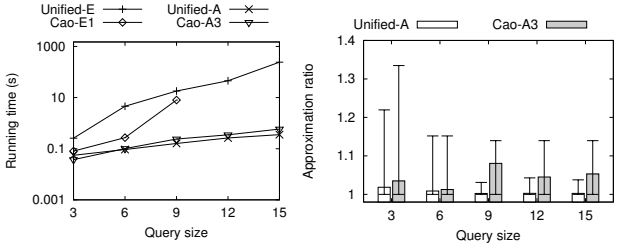
(6) $cost_{MaxMax2}$**.** The results for $cost_{MaxMax2}$ on the datasets GN and Web are shown in Figure 28 and Figure 29, respectively, which are similar to that on the dataset Hotel.

(7) $cost_{Max}$**.** The results for $cost_{Max}$ on the datasets GN and Web are shown in Figure 30, which is similar to that on the dataset Hotel. According to the results, both *Unified-E* and *Unified-A* run very fast, e.g. they ran less than 6 ms for all settings of $|q.\psi|$.



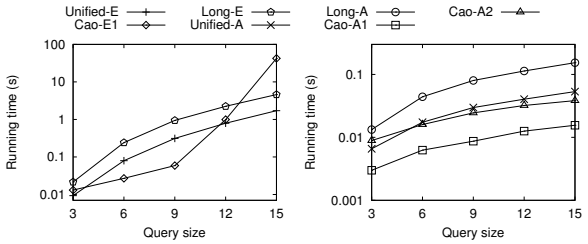
(a) Running time (b) Approximation ratio

Fig. 24: Effect of $|q.\psi|$ on $cost_{SumMax}$ (GN)

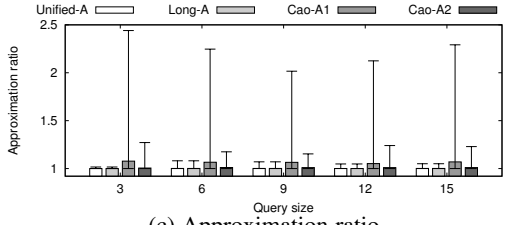


(a) Running time (b) Approximation ratio

Fig. 25: Effect of $|q.\psi|$ on $cost_{SumMax}$ (Web)

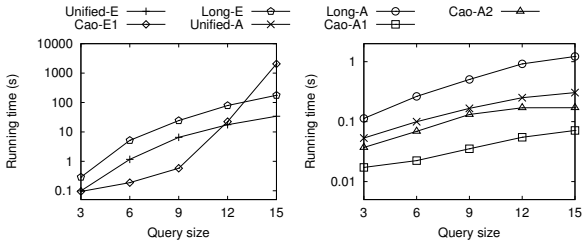


(a) Running time (Exact) (b) Running time (Appro.)

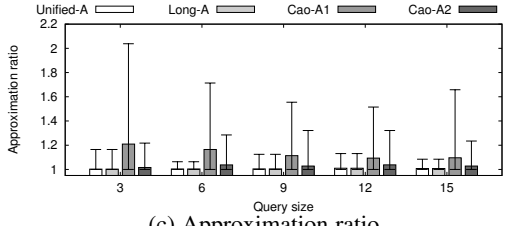


(c) Approximation ratio

Fig. 26: Effect of $|q.\psi|$ on $cost_{MaxMax}$ (GN)



(a) Running time (Exact) (b) Running time (Appro.)

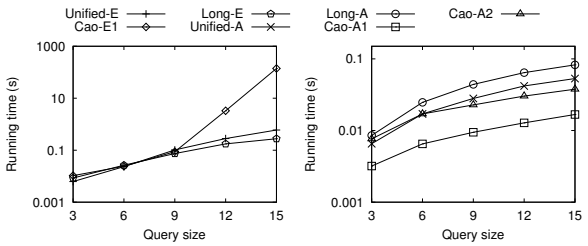


(c) Approximation ratio

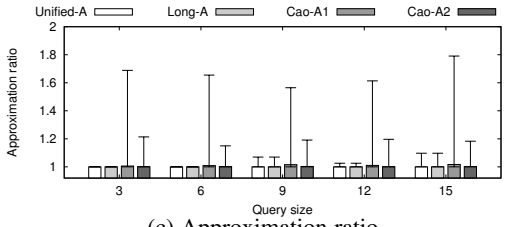
Fig. 27: Effect of $|q.\psi|$ on $cost_{MaxMax}$ (Web)

APPENDIX F
SCALABILITY TEST

(2) $cost_{MinMax2}$**.** The results for $cost_{MinMax2}$ are shown in Figure 31. According to Figure 31(a), *Unified-E* is faster and more scalable than *Cao-E1*, e.g., on a dataset with 6M objects, *Unified-E* ran for a couple of seconds while *Cao-E1* ran for more than 10 hours. Besides, similar to the case of $cost_{MinMax}$, *Unified-A* runs slightly slower than *Cao-A1*, but gives much better approximation ratio, e.g. the median of approximation ratios of *Unified-A* are 1 on all settings while that of *Cao-A1* are larger than 1.

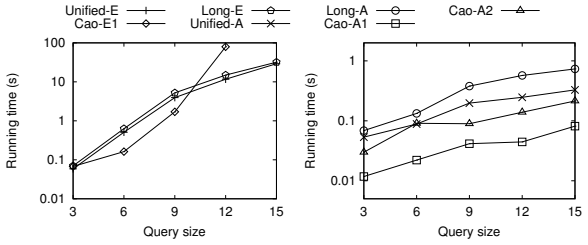


(a) Running time (Exact) (b) Running time (Appro.)

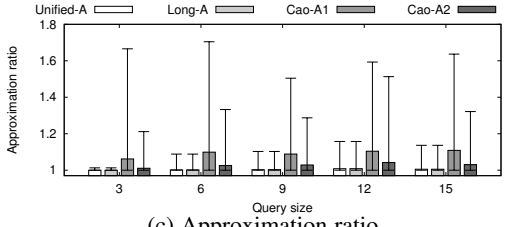


(c) Approximation ratio

Fig. 28: Effect of $|q.\psi|$ on $cost_{MaxMax2}$ (GN)

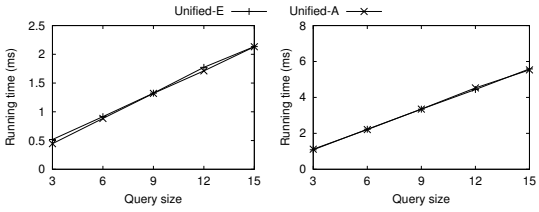


(a) Running time (Exact) (b) Running time (Appro.)



(c) Approximation ratio

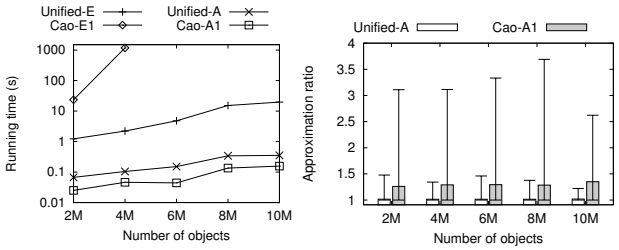
Fig. 29: Effect of $|q.\psi|$ on $cost_{MaxMax2}$ (Web)



(a) GN (b) Web

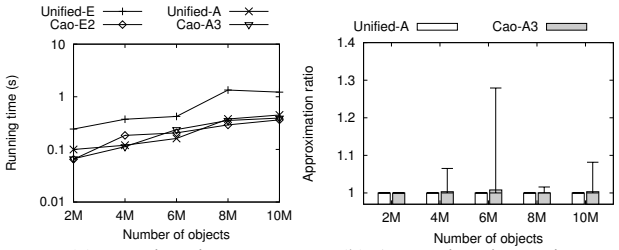
Fig. 30: Effect of $|q.\psi|$ on $cost_{Max}$

(3) $cost_{Sum}$**.** The results for $cost_{Sum}$ are shown in Figure 32. According to Figure 32(a), *Unified-E* is very scalable when the number of objects is large, e.g., it ran slightly longer than 1 second on a dataset with 10M objects. Besides, we noticed that *Cao-E2* has a very good performance and it even runs as fast as the approximation algorithms. The reason could be as follows. With the number of objects grows, the number of relevant objects



(a) Running time (b) Approximation ratio

Fig. 31: Scalability test on $cost_{MinMax2}$

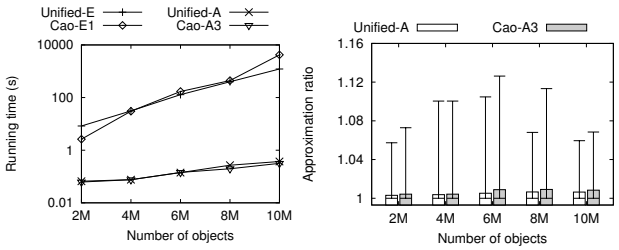


(a) Running time (b) Approximation ratio

Fig. 32: Scalability test on $cost_{Sum}$

becomes large. Both approximate algorithms have to re-compute the ratio for the remaining nodes in the heap and re-organize the heap after picking each object, whose cost becomes expensive when the number of relevant objects is large. In contrast, *Cao-E2* maintains a heap structure though, it does not have to re-examine the nodes after processing a node. *Unified-A* has similar running times as *Cao-A3* but gives better approximation ratios than *Cao-A3* (Figure 32(b)). Specifically, *Unified-A* can achieve near-to-optimal approximation ratios on all setting while *Cao-A3* has its largest approximation ratios up to 1.279.

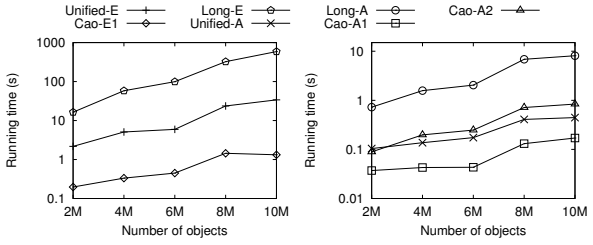
(4) $cost_{SumMax}$**.** Same as the experiments of varying $|o.\psi|$ for $cost_{SumMax}$, we used the setting of $|q.\psi| = 8$ for the scalability test experiments for $cost_{SumMax}$ particularly. The results for $cost_{SumMax}$ are shown in Figure 33. According to Figure 33(a), *Unified-E* and *Cao-E1* have similar running times and *Unified-A* and *Cao-A3* also have similar running times, but *Unified-A* gives a better approximation ratio than *Cao-A3* (Figure 33(b)).



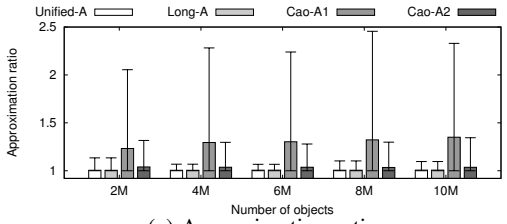
(a) Running time (b) Approximation ratio

Fig. 33: Scalability test on $cost_{SumMax}$

(5) $cost_{MaxMax}$**.** The results for $cost_{MaxMax}$ are shown in Figure 34. According to Figure 34(a), *Unified-E* runs faster than *Long-E* but slower than *Cao-E1*. According to Figure 34(b) and (c), *Unified-A* runs faster than *Long-A* and *Cao-A2* and slower than *Cao-A1*, and *Unified-A* is one of the two algorithms (the other is *Long-A* which runs slower than *Unified-A* by about one order of magnitude) which give the best approximation ratios. Specifically, the largest approximation ratios of *Unified-A* is only 1.134, which is small, while that of *Cao-A1* and *Cao-A2* are 2.456 and 1.345, respectively.



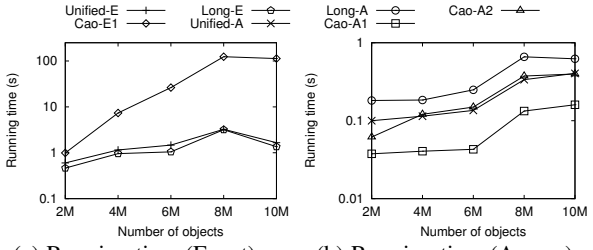
(a) Running time (Exact) (b) Running time (Appro.)



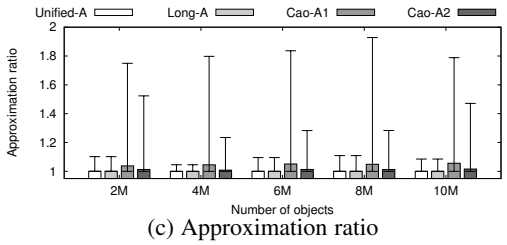
(c) Approximation ratio

Fig. 34: Scalability test on $cost_{MaxMax}$

(6) $cost_{MaxMax2}$**.** The results for $cost_{MaxMax2}$ are shown in Figure 35. According to Figure 35(a), *Unified-E* runs similarly fast as *Long-E*, and both of them run faster than *Cao-E1*. According to Figure 35(b) and (c), *Unified-A* has similar running times with *Cao-A2*, both of them run faster than *Long-A* and slower than *Cao-A1*, and *Unified-A* is one of the two algorithms (the other is *Long-A* which runs slower than *Unified-A*) which give the best approximation ratios. Specifically, the largest approximation ratios of *Unified-A* is only 1.109, which is small, while that of *Cao-A1* and *Cao-A2* are 1.928 and 1.524, respectively.



(a) Running time (Exact) (b) Running time (Appro.)



(c) Approximation ratio

Fig. 35: Scalability test on $cost_{MaxMax2}$

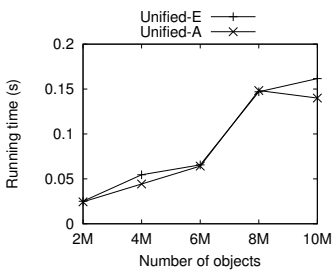


Fig. 36: Scalability test on $cost_{Max}$

(7) $cost_{Max}$**.** The results for $cost_{Max}$ are shown in Figure 36. According to the results, both *Unified-E* and *Unified-A* runs very fast, e.g. they ran within 1 second on a dataset with 10M objects.